LIMITED PROCESSOR SHARING QUEUES AND MULTI-SERVER QUEUES

A Thesis
Presented to
The Academic Faculty

by

Jiheng Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2009

LIMITED PROCESSOR SHARING QUEUES AND MULTI-SERVER QUEUES

Approved by:

Jim Dai, advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Bert Zwart, advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Hayriye Ayhan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Ton Dieker
School of Industrial and Systems Engineering
Georgia Institute of Technology

Amy Ward
Marshall School of Business
University of Southern California

Date Approved: 2 July 2009

ACKNOWLEDGEMENTS

I am deeply grateful to all people who have helped and inspired me during my doctoral study. None of my research would be possible without their help.

In the first place I would like to record my gratitude to my advisors, Dr. Jim Dai and Dr. Bert Zwart for their supervision, advice, and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. Their knowledge, insights and rigorous attitude toward research is a great gift to me, which I will carry with me during my career. They have been great mentors and tremendous source of motivation.

I am grateful to Dr. Christian Gromoll from the department of mathematics at University of Virginia. His pioneering work on the measure valued process has inspired me and is a foundation for my thesis work. In particular, I appreciate several notes he wrote me that suggested a nice method to rigorously prove Lemma 10.6.

I would like to thank Dr. Hayriye Ayhan, Dr. Ton Dieker and Dr. Amy Ward for serving on my thesis committee. In particular, I am especially thankful to Dr. Ton Dieker, for his role in making the presentation of this dissertation clearer.

Several papers have been produced based on the thesis, and are submitted or accepted by research journals. I am very grateful to anonymous referees, whose comments and suggestions have greatly improved this thesis.

I am also grateful to my Alma maters, Nanjing University and the Ohio State University, where I obtained my bachelor's and master's degree in mathematics, respectively. My learning experience there has built a solid foundation for my research.

The head of our graduate studies program, Dr. Gary Parker, has helped tremendously during my study at Georgia Tech. As a international student in the United

States, I got several visa problems during my trips outside of the country. Dr. Parker was always there and did the best of his effort in a timely fashion to help me out. Without his effort, I probably could not come back and finish my study. This is just one of many miracles he performed throughout my Ph.D. studies.

I am very grateful to have many friends here, without them my life would not have been nearly enjoyable and this work would not have been possible. Among them are Jieyun Zhou, Xuelei Ni, Jung-Kyung Kim, Theologos Bountourelis, Andrei Prudius, Josh Reed, Brian Fralix, Zhaosong Lv, Yang Zhang, Sebastian Ubinas. I would like to thank Tiger Lei Wu and Hussein Eser Kirkizlar for pushing me to go to gym regularly for the past 5 years, and Yao-Hsuan Chen for bringing me to the mountains and rock climbing when I need a refresh.

I would like to thank my parents and my girlfriend Lulu Kang, who inevitably shared my quest, for their unconditional love, constant support and encouragement. This thesis is dedicated to them.

Thanks to H. Milton Stewart School of Industrial and System Engineering at Georgia Institute of Technology and NSF grant CMMI-0727400 for financial support.

TABLE OF CONTENTS

VOLUME II MULTI-SERVER QUEUES

LIST OF TABLES

LIST OF FIGURES

SUMMARY

We study two classes of stochastic systems, the limited processor sharing system and the multi-server system. They share the common feature that multiple jobs/customers are being processed simultaneously, which makes the study of them intrinsically difficult.

In the limited processor sharing system, a limited number of jobs can equally share a single server, and the excess ones wait in a first-in-first-out buffer. The model is mainly motivated by computer related applications, such as database servers and packet transmission over the Internet. This model is studied in the first part of the thesis.

The multi-server queue is mainly motivated by call centers, where each customer is handled by an agent. The number of customers being served at any time is limited by number of agents employed. Customers who can not be served upon arrival wait in a first-in-first-out buffer. This model is studied in the second part of the thesis.

LIMITED PROCESSOR SHARING QUEUES AND MULTI-SERVER QUEUES

VOLUME I

Limited Processor Sharing Queues

by

Jiheng Zhang

CHAPTER I

INTRODUCTION

1.1 Motivation

We consider the limited processor sharing queue (LPS) which is a generalization of the processor sharing (PS) queue. As inferred from the name, we limit the number of jobs that can share the server at any time by $K \geq 1$, instead of letting all the jobs share the server. The server is shared equally by those jobs in service, i.e., at any time each job in service is processed at a rate that is the reciprocal of the number of jobs in service. An arriving job immediately starts receiving service if there are less than K jobs in the server when it arrives; otherwise it waits in the buffer. When the number of jobs in the server drops from K to $K-1$, the server immediately admits the longest waiting job from the buffer if there is any. A job leaves the system immediately after the server has fulfilled its service requirement. This is a quite general model since letting $K = \infty$ makes the system a PS queue and taking $K = 1$ reduces the system to a FCFS queue.

There is ample motivation to study this generalization. The PS model has been widely used in the analysis of computer systems, network servers and data transmission over the Internet. The PS discipline can be viewed as an idealization of time-sharing protocols in computer systems, as described in [37] and [49]. The advantage is that a big job will not block the whole system as in a FCFS queue. However, allowing too many jobs to time-share at once can lead to significant overhead (due to switching), hence reduce overall performance. This point has already been observed in early papers on operating systems [14, 6], as well as in more recent studies on Web server design [16, 34], and databases [29, 50]. So in applications, a sharing limit is

normally imposed, which results in the LPS model.

1.2 Related Literature

Despite its wide range of applications, there are only a few studies on the LPS queue. Avi-Itzhak and Halfin [3] propose an approximation for the mean response time assuming Poisson arrivals. A computational analysis based on matrix geometric methods is performed in Zhang and Lipsky [59, 60]. Some stochastic ordering results are derived in Nuyens and van de Weij [42]. No rigorous analysis for general job size distributions seems to be available.

Our study is carried out in a general setting, allowing the inter-arrival time and job sizes to have general distributions. Due to the general distribution of job sizes, the system is not Markovian. Since multiple jobs can be in service at the same time, the remaining job sizes for each of them become important in studying the dynamics of the system. For this purpose, we record all the remaining job sizes of all jobs in service using a measure $\mathcal{Z}(t)$ at any time t. For any Borel set $B \subset \mathbb{R}_+$, $\mathcal{Z}(t)(B)$ indicates the number of jobs in service with remaining job size belonging to B at that time. Similarly, we use a measure $\mathcal{Q}(t)$ to describe the state of the buffer, and $\mathcal{Q}(t)(B)$ indicates the number of jobs in the buffer with job size belonging to B. The descriptor $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$, which takes values in the space of two dimensional vectors of Borel measures, contains a wealth of information. All the usual performance processes can be recovered from it. For example, the total number of jobs in the server is $Z(\cdot) = \mathcal{Z}(\cdot)(\mathbb{R}_+)$, which can be written as

$$Z(\cdot) = \langle 1, \mathcal{Z}(\cdot) \rangle,$$

where the operator $\langle f, \mu \rangle$ in the above denotes the integration of function f against the measure μ. The workload $W(\cdot)$ in the system can be recovered by

$$W(\cdot) = \langle \chi, \mathcal{Q}(\cdot) + \mathcal{Z}(\cdot) \rangle,$$

where χ denote the identity function on $\mathbb{R}_+$. In fact, the measure valued descriptor contains all the information needed to describe the dynamics of the LPS system. More details will be discussed when we give a detailed model description in Chapter 2.

The framework of using measure-valued process has been successfully applied to study models where multiple jobs are processed at the same time. Existing works include Gromoll, Puha and Williams [25], Gromoll and Kruk [24] and Gromoll, Robert and Zwart [26]. However, in most of these works, buffers are not modeled because a job immediately starts service upon arrival. The only exception is Doytchinov, Lehoczky and Shreve [15]; in their model only one job is processed at a time and the buffer dynamics are described by a measure-valued process. As will be explained in the next two paragraphs, the existence of the buffer (due to the sharing limit) creates a big challenge in our study of fluid models and the corresponding fluid limits. One major difficulty is that the stochastic process of jobs entering the service is not as simple as the arrival process. When the system size is below K, that process is the same as the arrival process; when the system size is equal to K, the amount of jobs that enter the service in the next infinitely small amount of time equals the minimum of the arrivals and departures in that time period; when the system size is above K, the process of jobs entering the service equals the departure process. In short, the input into the server depends on the state of the system. We design a set of system dynamic equations (2.5) and (2.6) involving both the server $\mathcal{Z}(\cdot)$ and the buffer $\mathcal{Q}(\cdot)$. They are powerful enough to capture the complex dynamics, and yet simple enough to perform rigorous analysis.

1.3 Overview

The ultimate goal is to obtain insightful performance evaluation. No analytic tool is available to date that is capable of achieving the goal. Since the model we consider is a generalization of the $G/GI/1$ PS queue, and exact performance analysis of that

model seems intractable, our research focuses on obtaining approximations of the various performance quantities via limit theorems.

In order to develop tractable approximations for performance measures of the LPS queues, we first study the underlying stochastic processes in the heavy traffic regime, an asymptotic regime where the system becomes critically loaded and the sharing limit K becomes large. To study such a complicated stochastic system, we introduce a deterministic fluid model. This model is given by a set of *fluid dynamic equations*, which are deterministic analogs of the stochastic ones. After establishing several fundamental properties, such as existence and uniqueness of fluid model solutions, we show that the fluid model solution converges to an equilibrium state *uniformly* for all initial states within a compact set (c.f. Theorem 3.3). We next show that this fluid model arises as the limit of fluid scaled systems of LPS queues. This part of the analysis applies to a variety of regimes, such as lightly loaded, critically loaded and overloaded systems. The fluid model and fluid approximations are of independent interest. More importantly, they pave the way to the study of heavy traffic limits of the diffusion scaled processes

$$\left(\hat{\mathcal{Q}}^r(\cdot), \hat{\mathcal{Z}}^r(\cdot)\right) = \left(\frac{1}{r}\mathcal{Q}^r(r^2\cdot), \frac{1}{r}\mathcal{Z}^r(r^2\cdot)\right)$$

as r goes to infinity. As it has been shown in Williams [57], a key step to obtain a diffusion limit in heavy traffic is to establish a *state space collapse* (SSC) result. In our setting, the SSC means that the diffusion scaled measure valued process, which is an infinite dimensional object, is close to a deterministic functional (c.f. Definition 3.4) of the diffusion scaled workload process, which is one-dimensional. It is well known that the diffusion limit of the workload process is a one-dimensional reflected Brownian motion (RBM) for any non-idling service policy, including the LPS one. A major objective is to show that our measure valued diffusion limit is a deterministic functional of the one-dimensional RBM (c.f. Theorem 5.1). As a corollary, the diffusion scaled total job size process converges in distribution to a

piecewise linear RBM. Since the diffusion limit, which is denoted by $(\mathcal{Q}^*(\cdot), \mathcal{Z}^*(\cdot))$, is a functional of RBM, it is possible to obtain various steady state properties. As we show in (6.3), the steady state limit $(\mathcal{Q}^*(\infty), \mathcal{Z}^*(\infty))$ has a rather explicit distribution.

The general framework of the above analysis has been developed in the literature, e.g. [57, 7]. In particular, a sequence of papers [25, 45, 23] apply the same framework to study the processor sharing queue. Instead of being a straightforward extension of the processor sharing queue, the study of LPS queue is quite challenging and requires innovative ideas and techniques, mainly due to the sharing limit K. First, the fluid model involves a complicated functional equation (after some mathematical derivations including a time change) in our analysis:

$$x(u) = h(u) + \int_0^u (x(u-v) - K)^+ dF(v) + \rho \int_0^u (x(u-v) \wedge K) dF_e(v), \qquad (1.1)$$

where ρ is the traffic intensity, F is the job size distribution and F_e is the equilibrium distribution of F (c.f. see Section 3.3 for background and notation). In the special case of the standard PS queue, $K = \infty$ and this equation reduces to a standard renewal equation. Existence and uniqueness of the solution to a renewal equation is already known. But in general, this is not the case, necessitating new methods. The counterpart of proving uniform convergence of fluid model solution to the equilibrium for standard PS queues has been carried out by Puha and Williams [45]. Standard PS queues are relatively tractable since (1.1) is essentially a renewal equation, so the key renewal theorem can be applied. However, it requires the development of new tools to deal with general K in the functional equation (1.1). We have developed general tools to study the functional equation, which are new to the best of our knowledge. Second, we need to prove the tightness property in order to establish the fluid and diffusion approximation. Due to the fact that the process of jobs entering the service is not as simple as the arrival process in the PS queue (due to sharing level $K < \infty$ in LPS queue), we need new techniques to control the internal process that jobs move from the buffer to the server. Third, since the object of study is a random measure

instead of real numbers, we need a certain type of Borel-Cantelli estimates for the difference between the empirical distribution and the limit distribution. Again, due to the complexity caused by the sharing limit K, we need a more general version, which we proved from scratch. In fact, the sharing limit K caused quite substantial difficulties throughout the thesis.

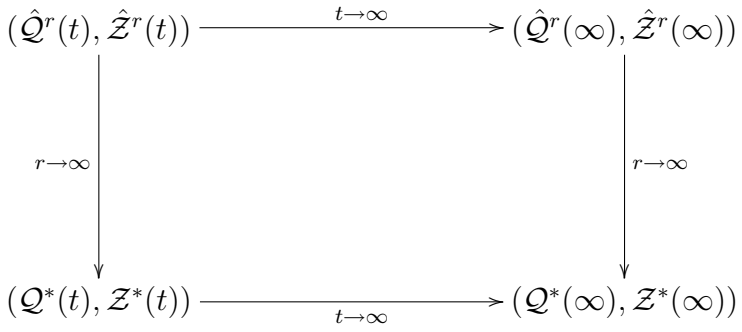

Figure 1: The interchange of heavy traffic and steady state limits.

The original stochastic measure-valued process is regenerative, and will converge weakly as $t \to \infty$ to a steady state (c.f. Theorem 6.1). However, no explicit solution for the stationary distribution seems available. So we establish the interchange of heavy traffic and steady state limits as depicted in Figure 1.3. The main idea is to couple the measure valued process for the LPS queue with its corresponding stationary version. This helps obtain a version of classical coupling inequality, cf. (6.7). The interchange can be established after we prove the uniform convergence of the upper bound for the coupling inequality. It should be pointed out that the framework of using coupling to establish interchange of the heavy traffic and steady state limit works for all single buffer single server system in workload conserving disciplines. In particular, for the classical PS queue our technique works as well, and allows one to recover Grischeckin's [22] steady-state approximation from Gromoll's [23] process limit theorem. For networks, the interchange is more involved, cf. Budhiraja & Lee [9] and Gamarnik & Zeevi [19].

The validity of the interchange provides the necessary theoretical support for using

the tractable limit $(\mathcal{Q}^*(\infty), \mathcal{Z}^*(\infty))$ as an approximation of the steady state of a given LPS queue. Section 6.2 demonstrates how to analyze performance quantities such as queue size, delay probability and response times via limit theorems. From a practical perspective, the main insights are the approximation formulas (6.27) and (6.28) for queue size delay probability, and (6.29)–(6.31) for response times. In particular, our results show that the following two-moment approximation of the queue size $E[X]$ is accurate:

$$\mathbb{E}[X] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{\rho}{1-\rho}(1 - d_p) + \frac{c_a^2 + c_s^2}{2} \frac{\rho}{1-\rho} d_p,$$

where $d_p = \rho^{\frac{1+c_s^2}{c_a^2+c_s^2}K}$. In the above display, ρ is the traffic intensity, c_a^2 and c_s^2 are coefficients of variation for the inter-arrival and job sizes. The value d_p can be interpreted as the approximation for the probability that a customer cannot get service immediately upon arrival. Interestingly, our approximation is consistent with the heuristic of Avi-Itzhak and Halfin [3] for the Poisson arrival process.

This part of the thesis is organized as follows. A detailed model description is given in Chapter 2. Chapter 3 investigates properties of the fluid models, including existence, uniqueness and uniform convergence to the equilibrium. Convergence of a fluid scaled sequence of systems to fluid model solution is studied in Chapter 4. Chapter 5 establish the state space collapse property and the diffusion limit. Several auxiliary tools and results are proved in the appendices. In Chapter 6, we study the steady state limit of the LPS queue, and establish the validity of interchanging the heavy traffic and steady limit of the LPS queue. The interchange provides the foundation for performance analysis later in that chapter.

The existence and uniqueness results in Chapter 3 and Chapter 4 have been published in the paper [63] with Jim Dai and Bert Zwart. The uniform convergence to the equilibrium in Chapter 3 and and Chapter 5 have been summarized in the technical report [62] with Jim Dai and Bert Zwart. Chapter 6 is published in the paper [64] with Bert Zwart.

1.4 Notation

The following notation will be used throughout. Let $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ denote the set of natural numbers, integers and real numbers respectively. Let $\mathbb{R}_+ = [0, \infty)$. For $a, b \in \mathbb{R}$, write a^+ for the positive part of a, $\lfloor a \rfloor$ for the integer part, $\lceil a \rceil$ for $\lfloor a \rfloor + 1$, $a \vee b$ for the maximum, and $a \wedge b$ for the minimum.

Let $\mathbf{M}$, $\mathbf{M}_1$ and $\mathbf{M}_2$ denote the set of all non-negative finite Borel measures on $\mathbb{R}$, $[0, \infty)$ and $(0, \infty)$, respectively. To simplify the notation, let us take the convention that for any Borel set $A \subset \mathbb{R}$, $\nu(A \cap (-\infty, 0)) = 0$ for any $\nu \in \mathbf{M}_1$ and $\nu(A \cap (-\infty, 0]) = 0$ for any $\nu \in \mathbf{M}_2$. Also, by this convention, $\mathbf{M}_2$ is embedded as a subspace of $\mathbf{M}_1$, which is also a subspace of $\mathbf{M}$. For $\nu_1, \nu_2 \in \mathbf{M}_1$, the Prohorov metric is defined to be

$$\mathbf{d}[\nu_1, \nu_2] = \inf \Big\{ \epsilon > 0 : \nu_1(A) \le \nu_2(A^\epsilon) + \epsilon \text{ and } \nu_2(A) \le \nu_1(A^\epsilon) + \epsilon \text{ for all closed Borel set } A \subset \mathbb{R}_+ \Big\},$$

where $A^\epsilon = \{b \in \mathbb{R}_+ : \inf_{a \in A} |a - b| < \epsilon\}$. This is the metric that induces the topology of weak convergence of finite Borel measures. (See Section 1.6 in [5] and Section4.3 in [25].) For any Borel measurable function $g : \mathbb{R}_+ \to \mathbb{R}$, the integration of this function with respect to the measure $\nu \in \mathbf{M}_1$ is denoted by $\langle g, \nu \rangle$.

Let $\mathbf{M}_1 \times \mathbf{M}_2$ denote the Cartesian product. There are a number of ways to define the metric on the product space. For convenience we define the metric to be the maximum of the Prohorov metric between each component. With a little abuse of notation, we still use $\mathbf{d}$ to denote this metric.

Let $(\mathbf{E}, \pi)$ be a general metric space. We consider the space $\mathbf{D}$ of all right-continuous $\mathbf{E}$-valued functions with finite left limits defined either on a finite interval $[0, T]$ or the infinite interval $[0, \infty)$. We refer to the space as $\mathbf{D}([0, T], \mathbf{E})$ or $\mathbf{D}([0, \infty), \mathbf{E})$ depending upon the function domain. The space $\mathbf{D}$ is also known as the space of *càdlàg* functions. For $g(\cdot), g'(\cdot) \in \mathbf{D}([0, T], \mathbf{E})$, the uniform metric is defined

as

$$\upsilon_T[g, g'] = \sup_{0\le t\le T} \pi[g(t), g'(t)]. \tag{1.2}$$

Another metric we will use is the following Skorohod J_1 metric,

$$\varrho_T[g, g'] = \inf_{f\in\Lambda_T} (\|f\|^\circ_T \vee \upsilon_T[g, g' \circ f]), \tag{1.3}$$

where $g\circ f(t) = g(f(t))$ for $t \ge 0$ and Λ_T is the set of strictly increasing and continuous mapping of $[0, T]$ onto itself and

$$\|f\|^\circ_T = \sup_{0\le s<t\le T} \big|\log \frac{f(t) - f(s)}{t - s}\big|.$$

If $g(\cdot)$ and $g'(\cdot)$ are in the space $\mathbf{D}([0, \infty), \mathbf{E})$, the Skorohod J_1 metric is defined as

$$\varrho[g, g'] = \int_0^\infty e^{-T}(\varrho_T[g, g'] \wedge 1)dT. \tag{1.4}$$

By saying convergence in the space $\mathbf{D}$, we mean the convergence under the Skorohod J_1 topology, which is the topology induced by the Skorohod J_1 metric [17].

We use "$\to$" to denote the convergence in a general metric space $(\mathbf{E}, \pi)$, and use "$\Rightarrow$" to denote the convergence in distribution of random variables taking value in the metric space $(\mathbf{E}, \pi)$.

CHAPTER II

MODEL DESCRIPTION

We first rigorously introduce the mathematical model underline the LPS queue in this chapter. Consider a $G/GI/1$ queue operated under the limited processor sharing policy, with the sharing limit equal to K. We use $Q(t)$, $Z(t)$, and $X(t)$ to denote the number of jobs in the buffer, number of jobs in service, and the total number of jobs in the system at time t, respectively. Thus,

$$X(t) = Q(t) + Z(t) \quad \text{for } t \geq 0. \tag{2.1}$$

The system is allowed to be non-empty initially, i.e. $X(0) > 0$. We index jobs by $i = -X(0)+1, -X(0)+2, \ldots, 0, 1, \ldots$. The first $X(0)$ jobs are initially in the system, with jobs $i = -X(0)+1, \ldots, -Q(0)$ in service and jobs $i = -Q(0)+1, \ldots 0$ waiting in the buffer. Jobs arrived after time 0 are indexed by $i = 1, 2 \ldots$. Let $E(t)$ denote the number of jobs that arrive to the buffer during time interval $(0, t]$, for all $t \geq 0$. According to the policy, a job may have to wait for a certain amount of time after arrival to get service. Let w_i denote the waiting time, and U_i denote the arrival time of the ith job for all $i > -X(0)$. By convention, $U_i = 0$ for $i < 0$, and $w_i = 0$ for $i \leq -Q(0)$. Let

$$\tau_i = U_i + w_i, \quad i > -X(0).$$

The quantity τ_i can be viewed as the time that the ith job starts service. We use v_i to denote the job size of the ith job for all $i > -Q(0)$. We assume that $\{v_i\}_{i=-\infty}^{\infty}$ is a sequence of i.i.d. random variables with distribution F. For jobs with index $-X(0) < i \leq -Q(0)$, i.e. the first $Z(0)$ jobs that are in service initially in service, we use $\tilde{v}_i$ to denote the remaining job size of the job. The sequence $\{\tilde{v}_i\}_{i=-\infty}^{0}$ is allowed

to be general. We call $\{E(\cdot), \{v_i\}_{i=1}^{\infty}\}$ the stochastic primitives of the system, and $\{Z(0), Q(0), \{v_i\}_{i=-\infty}^{0}, \{\tilde{v}_i\}_{i=-\infty}^{0}\}$ the initial conditions of the system.

Now we introduce a measure-valued state descriptor $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$, which describes the evolution of the system with given initial conditions and stochastic primitives. Let $\mathcal{Q}(\cdot)$ and $\mathcal{Z}(\cdot)$ be $\mathbf{M}_1$-valued and $\mathbf{M}_2$-valued stochastic processes, respectively. For any Borel set $A \subset [0, \infty)$, $\mathcal{Q}(t)(A)$ denotes the total number of jobs in buffer whose job size belongs to A; for any Borel set $A \subset (0, \infty)$, $\mathcal{Z}(t)(A)$ denotes the total number of jobs in service whose residual job size belongs to set A. Note that here we distinguish the spaces for buffer and server descriptors. The reason is that we allow jobs with size 0 to arrive and wait in the buffer. However, a job in service will immediately leave the system once its remaining service time becomes 0. So no job in service can have zero remaining service time. It is clear that we have the following relationship

$$Q(t) = \langle 1, \mathcal{Q}(t) \rangle, \quad Z(t) = \langle 1, \mathcal{Z}(t) \rangle.$$

Define the *cumulative service amount* up to time t by

$$S(t) = \int_0^t \psi(Z(\tau)) d\tau, \tag{2.2}$$

where $\psi(x) = 1/x$ if $x > 0$ and $\psi(x) = 0$ if $x = 0$. A job will have received a cumulative amount of processing time

$$S(s, t) = \int_s^t \psi(Z(\tau)) d\tau$$

during time interval $[s, t]$ if it is in service in this time period. Let

$$B(t) = E(t) - Q(t). \tag{2.3}$$

Note that at time $t \geq 0$, $B(t)$ is the index of the last job who has entered into service by time t. Thus

$$B(s, t) = B(t) - B(s) \tag{2.4}$$

represents the number of jobs which have left the buffer and entered the server during time interval $(s, t]$. Using the notation introduced in this section, the state descriptor can be written as

$$\mathcal{Q}(t)(A') = \sum_{i=B(t)+1}^{E(t)} \delta_{v_i}(A') \tag{2.5}$$

$$\mathcal{Z}(t)(A) = \sum_{i=-X(0)+1}^{-Q(0)} \delta_{\tilde{v}_i}(A + S(t)) + \sum_{i=-Q(0)+1}^{B(t)} \delta_{v_i}(A + S(\tau_i, t)), \tag{2.6}$$

for any Borel sets $A' \subseteq [0, \infty)$ and $A \subseteq (0, \infty)$ and $t \geq 0$, where δ_a denotes the Dirac measure of point a on $\mathbb{R}$ and $A + y = \{a + y : a \in A\}$. Due to the LPS policy, the sharing limit K must be enforced at any time t,

$$Q(t) = (X(t) - K)^+, \tag{2.7}$$

$$Z(t) = (X(t) \wedge K). \tag{2.8}$$

We call (2.5) and (2.6) the *stochastic dynamic equations* and (2.7) and (2.8) policy constraints.

For $t \geq 0$, the workload of the system $W(t)$ is defined to be the amount of time that the server remains busy *if* no more arrivals are allowed into the system at time t. Using the state descriptor $(\mathcal{Q}, \mathcal{Z})$, we can recover the workload $W(t)$ at time $t \geq 0$ by

$$W(t) = \langle \chi, \mathcal{Q}(t) + \mathcal{Z}(t) \rangle, \tag{2.9}$$

where χ denotes the identity function on $\mathbb{R}$.

CHAPTER III

FLUID MODEL AND ITS PROPERTIES

To study the LPS queue introduced in the previous chapter, we first introduce a corresponding measure-valued fluid model. For this fluid model, we establish several fundamental properties, such as existence and uniqueness of fluid model solutions. We also characterize the equilibrium state of the fluid model and establish the uniform convergence of fluid model solutions to the equilibrium. It will be shown in Chapter 4 that fluid model arises as the limit of fluid scaled systems of LPS queues. Our analysis applies to a variety of regimes, such as lightly loaded, critically loaded and overloaded systems. The approximation properties of fluid model will help establish the heavy traffic limit theorems.

A difficulty in our study is that the fluid model involves a complicated functional equation, (3.23), after some mathematical derivations including a time change. In the special case of the standard PS queue, $K = \infty$ and this equation reduces to a standard renewal equation. Existence and uniqueness of the solution to a renewal equation is already known. In our case, K is finite, necessitating new methods. To establish the uniform convergence of fluid model to the equilibrium, equation (3.23) also plays a key role. The challenge there is to obtain a version of the renewal theory with uniform convergence base on this equation. The methodology we used to study the equation and it's application to the fluid model is new to the best of our knowledge.

3.1 Fluid Model

In this section, we propose a fluid analogue of the LPS system. Given a measure-valued process $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}_1 \times \mathbf{M}_2)$, for $t \geq 0$, let

$$\bar{Q}(t) = \langle 1, \bar{\mathcal{Q}}(t) \rangle, \tag{3.1}$$

$$\bar{Z}(t) = \langle 1, \bar{\mathcal{Z}}(t) \rangle, \tag{3.2}$$

$$\bar{X}(t) = \bar{Q}(t) + \bar{Z}(t), \tag{3.3}$$

$$\bar{B}(t) = \lambda t - \bar{Q}(t), \tag{3.4}$$

$$\bar{D}(t) = \lambda t + \bar{X}(0) - \bar{X}(t), \tag{3.5}$$

where λ is a positive constant which is interpreted as the arrival rate. These quantities are the fluid analogues of $Q(t), Z(t), B(t), D(t)$ and $X(t)$ in the stochastic model. Define the *fluid cumulative service amount* up to time t by

$$\bar{S}(t) = \int_0^t \phi_\rho(\bar{Z}(\tau)) d\tau, \tag{3.6}$$

where $\phi_\rho(x) = 1/x$ for all $x, \rho > 0$ and

$$\phi_\rho(0) = \begin{cases} \infty & \rho \in (0,1], \\ 0 & \rho \in (1,\infty). \end{cases} \tag{3.7}$$

And for $0 \leq s \leq t$, denote

$$\bar{S}(s,t) = \int_s^t \phi_\rho(\bar{Z}(\tau)) d\tau. \tag{3.8}$$

This is how the fluid cumulative service amount is defined, and it turns out that this definition serves the purpose of studying the fluid model very well. Here we give some intuitive explanation of why using the function ϕ_ρ instead of ψ in (2.2). In the corresponding stochastic process, when there is no job in the system, the server idles, implying $\psi(0) = 0$. In the fluid model with $\rho \leq 1$, intuitively, the amount of fluid in service $\bar{Z}(\cdot)$ will stay at zero once it reaches zero. Since fluids flow in at a

constant rate λ, the server, instead of idling, actually finishes service immediately when an infinitesimal amount of fluid enters service. So very naturally, $\psi_\rho(0) = \infty$ when $\rho \leq 1$. However, when $\rho > 1$, intuitively, the queue size should grow if starts at zero. To rule out the solution $z(\cdot) \equiv 0$, we define $\psi_\rho(0) = 0$. Note that the definitions of fluid model solutions for the standard PS queue also depend on the load (cf. [25, 44]).

Let ν be the probability measure associated with the job size distribution F. We call ν the job size measure. An element $(\xi, \mu) \in \mathbf{M}_1 \times \mathbf{M}_2$ is called a *valid initial condition* if

$$\xi = (\langle 1, \xi \rangle + \langle 1, \mu \rangle - K)^+ \nu,$$
$$\langle 1, \mu \rangle = (\langle 1, \xi + \mu \rangle) \wedge K.$$

Roughly speaking, validity of an initial state means that the initial state is consistent with the limited sharing policy; initial waiting jobs have the same service distribution as arriving jobs. Denote $\mathscr{I}$ the set of all valid initial conditions.

We now introduce the following *fluid dynamic equations*, which are analogous to (2.5) and (2.6). For all $A_y = (y, \infty)$, $y \geq 0$,

$$\bar{\mathcal{Q}}(t)(A_y) = \xi(A_y) + \Big(\bar{Q}(t) - \bar{Q}(0)\Big)\nu(A_y), \tag{3.9}$$
$$\bar{\mathcal{Z}}(t)(A_y) = \mu(A_y + \bar{S}(t)) + \int_0^t \nu(A_y + \bar{S}(s,t))d\bar{B}(s), \tag{3.10}$$

where $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$, $\bar{X}(\cdot)$, $\bar{B}(\cdot)$ and $\bar{S}(\cdot)$ are defined in (3.1)–(3.8). They are subject to the following constraints:

$$\bar{B}(\cdot) \text{ is non-decreasing,} \tag{3.11}$$
$$\bar{Q}(t) = (\bar{X}(t) - K)^+, \tag{3.12}$$
$$\bar{Z}(t) = (\bar{X}(t) \wedge K). \tag{3.13}$$

The above equations define a fluid model, which we denote by the triple (K, λ, ν).

Denote $\beta = \langle \chi, \nu \rangle$ the mean of the job size, and define

$$\rho = \lambda\beta$$

to be the *traffic intensity* of the fluid model.

Definition 3.1. $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}_1 \times \mathbf{M}_2)$ *is a solution to the fluid model* (K, λ, ν) *with a valid initial state* (ξ, μ) *if it satisfies the fluid dynamic equations* (3.9) *and* (3.10)*, subject to the constraints* (3.11)–(3.13)*.*

Similar to the stochastic model, the fluid workload $\bar{W}(t)$ at any time $t > 0$ is defined as

$$\bar{W}(t) = \langle \chi, \bar{\mathcal{Q}}(t) + \bar{\mathcal{Z}}(t) \rangle. \tag{3.14}$$

3.2 Existence and Uniqueness of the Fluid Model Solution

We first present several key properties of our fluid model solution, including the existence and uniqueness, the workload conserving property, and criteria for stability. The results are established by considering different cases, depending on whether initial condition is zero or not, and whether traffic intensity is bigger than one or not. We now state these results in the following theorems and property. The proofs can be find at the end of this section.

The following theorem establishes the existence and uniqueness for a fluid model solution.

Theorem 3.1. *Assume that the job size measure* ν *satisfies*

$$\langle \chi, \nu \rangle < \infty, \tag{3.15}$$

$$\nu(\{0\}) = 0. \tag{3.16}$$

There exists a unique solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *to the fluid model* (K, λ, ν) *with initial condition* $(\xi, \mu) \in \mathscr{I}$*.*

We have the following workload conserving property for any fluid model solution.

Proposition 3.1. *Assume that the job size measure ν satisfies* (3.15) *and* (3.16). *The fluid workload $\bar{W}(\cdot)$ of any solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ to the fluid model (K, λ, ν) with initial condition $(\xi, \mu) \in \mathscr{I}$ satisfies*

$$\bar{W}(t) = (\langle \chi, \xi + \mu \rangle + (\rho - 1)t)^+ \quad \textit{for all } t \geq 0.$$

We now turn to stability properties of our fluid model. Although the results are intuitively clear, the stability properties of fluid model solutions require formal proof in the measure-valued setup. The following definitions are analogous to the standard fluid model as in Dai [11, 12].

Definition 3.2. *A fluid model (λ, K, ν) is* weakly stable *if any fluid model solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ with initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$ satisfies $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$ for all $t \geq 0$.*

A fluid model (λ, K, ν) is stable *if for any initial condition $(\xi, \mu) \in \mathscr{I}$ satisfying $0 < w = \langle \chi, \xi + \mu \rangle < \infty$, there exists a finite time δ (only depending on w) such that any fluid model solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ with this initial condition satisfies $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$ for all $t \geq \delta$.*

Theorem 3.2. *Assume that the job size measure ν satisfies* (3.15) *and* (3.16). *A fluid model (λ, K, ν) is weakly stable if the traffic intensity $\rho \leq 1$; it is stable if the traffic intensity $\rho < 1$.*

Recall that $A_y = (y, \infty)$ for all $y \geq 0$. Since the fluid amount of jobs in service $\bar{Z}(t) = \bar{\mathcal{Z}}(A_0)$ for all $t \geq 0$, according to (3.10) in Definition 3.1, we have

$$\bar{Z}(t) = \mu(A_{\bar{S}(t)}) + \int_0^t [1 - F(\bar{S}(s,t))] d\bar{B}(s). \tag{3.17}$$

To further analyze the fluid model, we need to distinguish between different cases. We first consider the case where the initial condition is non-zero. In this case, there

exists a non-trivial interval on which the amount of fluid in service never reaches zero. So we can do a time change to obtain the equation (3.23), which is the key equation in our analysis. Through this analysis, we can characterize the fluid model solution on a small interval. We then use the "restarting" lemma (Lemma 3.2) to extend the result to a larger interval. After that case, we consider the case where the initial condition is zero and traffic intensity $\rho \leq 1$. Basically, we show that the fluid model solution will stay at zero. Finally, we study the case with zero initial condition and $\rho > 1$. Briefly speaking, the fluid model solution will grow "linearly" in this case.

3.2.1 Starting with a Non-zero Valid Initial Condition

If the valid initial condition $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$, then $Z(0) = \langle 1, \mu \rangle > 0$. Let

$$t^* = \inf\{s > 0 : \bar{Z}(s) = 0\}. \tag{3.18}$$

Since $\bar{Z}(0) > 0$, by right-continuity of $\bar{Z}(\cdot)$ we have $t^* > 0$. The following calculations will be performed on the interval $[0, t^*)$, where the function $\bar{S}(\cdot)$ as defined in (3.6) has an inverse, which is denoted by $\bar{T}(\cdot)$. By the inverse function theorem,

$$\bar{T}'(v) = \bar{Z}(\bar{T}(v)). \tag{3.19}$$

Perform the change of variables $u = \bar{S}(t)$ and $v = \bar{S}(s)$ to (3.10), we get

$$\begin{aligned}\bar{Z}(\bar{T}(u)) = \xi(A_u) + \lambda \int_0^u [1 - F(u-v)]\bar{Z}(T(v))dv \\ - \int_0^u [1 - F(u-v)]d\bar{Q}(\bar{T}(v)).\end{aligned}$$

Through the change of variable $v \leftarrow u - v$ and integration by parts, we obtain

$$\begin{aligned}\bar{Z}(\bar{T}(u)) = \xi(A_u) + \lambda\beta \int_0^u \bar{Z}(T(u-v))dF_e(v) - [1 - F(0)]\bar{Q}(\bar{T}(u)) \\ + [1 - F(u)]\bar{Q}(0) + \int_0^u \bar{Q}(\bar{T}(u-v))dF(v),\end{aligned}$$

where F_e is the equilibrium distribution of F which can be written as $F_e(x) = \frac{1}{\beta}\int_0^x [1 - F(y)]dy$. By condition (3.16), $F(0) = 0$. Now we obtain the key relationship

$$\begin{aligned}\bar{Q}(\bar{T}(u)) + \bar{Z}(\bar{T}(u)) = \xi(A_u) + \mu(A_u) + \int_0^u \bar{Q}(\bar{T}(u-v))dF(v) \\ + \rho\int_0^u \bar{Z}(\bar{T}(u-v))dF_e(v)\end{aligned} \tag{3.20}$$

for all $0 \le u < u^* = \bar{S}(t^*)$. To simplify notation, denote

$$h(u) = \xi(A_u) + \mu(A_u), \tag{3.21}$$

$$x(u) = q(u) + z(u), \tag{3.22}$$

where $q(u) = \bar{Q}(\bar{T}(u))$ and $z(u) = \bar{Z}(\bar{T}(u))$. By (3.12) and (3.13), the above equation can be written as

$$x(u) = h(u) + \int_0^u (x(u-v) - K)^+ dF(v) + \rho\int_0^u (x(u-v) \wedge K)dF_e(v). \tag{3.23}$$

This equation would simplify to a renewal equation if $K = \infty$ or $K = 0$, which corresponds to the PS queue and FIFO queue respectively. In fact, although the fluid model in earlier works on the PS queue [25, 23] or related models [24, 26] is defined in a different way, the mathematical analysis is essentially focused on equation (3.23) with $K = \infty$. In our case where K is finite, equation (3.23) is not a renewal equation anymore. Resolving the substantial technical difficulties that arise when this is not the case is the main task in studying the fluid model for the LPS queue.

We provide a general tool to study the integral equation (3.23) in Appendix A. The tool represents one of our major technical contributions of the thesis. Lemma A.1 there requires even weaker conditions than we need, which may be useful in future work. In our setting, condition (3.16) and the definition of $h(\cdot)$ in (3.21) imply that all the conditions needed in that lemma are satisfied. Building on this lemma, we establish the existence and uniqueness of fluid model solutions on a small interval.

Lemma 3.1. *Assume* (3.15) *and* (3.16). *For any non-zero initial condition* $(\xi, \mu) \in \mathscr{I}$, *there exists a* $t' > 0$ *such that the fluid model* (K, λ, ν) *has a unique solution* $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *on* $[0, t']$ *satisfying the initial condition and*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) \neq (\mathbf{0}, \mathbf{0}) \quad \textit{for all } t \in [0, t').$$

Proof. Lemma A.1 establishes the uniqueness and existence of solution to (3.23) on the interval $[0, a]$, where a is positive and does not depend on initial condition. Let

$$a' = \inf\{u \leq a : x(u) = 0\}. \tag{3.24}$$

We have that $a' > 0$ since $x(\cdot)$ is right-continuous and the initial condition is non-zero. Now let

$$\bar{T}(u) = \int_0^u (x(v) \wedge K) dv.$$

It is clear that $\bar{T}(\cdot)$ is differentiable and strictly increasing on $[0, a']$. Let $\bar{S}(t)$ denote its inverse function, which is still differentiable and strictly increasing on $[0, a']$. Now define

$$\bar{X}(t) = x(\bar{S}(t))$$

and $\bar{Q}(t) = (\bar{X}(t) - K)^+$, $\bar{Z}(t) = \bar{X}(t) \wedge K$. Since $x(\cdot)$ is *càdlàg* and $\bar{T}(\cdot)$ is continuous, it is clear that $\bar{Q}(\cdot)$ is *càdlàg*. So are $\bar{Q}(\cdot)$ and $\bar{Z}(\cdot)$ as well. By the inverse function theorem,

$$\bar{S}'(t) = \frac{1}{T'(\bar{S}(t))} = \frac{1}{\bar{X}(t) \wedge K} = \frac{1}{\bar{Z}(t)}.$$

Since $x(\cdot)$ is a solution to (3.23) on the interval $[0, a']$, $\bar{Q}(\cdot)$ is a solution to (3.20) (and thus to (3.17)) on the interval $[0, t']$, where

$$t' = \bar{T}(a'). \tag{3.25}$$

Let $\bar{B}(t) = \lambda t - \bar{Q}(t)$ for all $t \in [0, t']$. Since (ξ, μ) is a valid initial condition, $\xi([0, u]) = (\langle 1, \xi + \mu \rangle - K)^+ F(u)$ and $\langle 1, \mu \rangle = \langle 1, \xi + \mu \rangle \wedge K$. Since $\mu \neq \mathbf{0}$, let $G(\cdot) = \mu([0, \cdot])/\langle 1, \mu \rangle$, which is a distribution function. Recall the definition of $h(\cdot)$

in (3.21), we have that $h(u) = (h(0) \wedge K)[1 - G(u)] + (h(0) - K)^+[1 - F(u)]$. Thus, it satisfies the conditions in Lemma A.2. So, by Lemma A.2, $\bar{B}(\bar{T}(u))$ is non-decreasing on the interval $[0, a']$. Thus, $\bar{B}(t)$ is non-decreasing on the interval $[0, t']$ due to the fact that $\bar{T}(u)$ is strictly increasing on $[0, a']$. Define

$$\bar{\mathcal{Q}}(t)(A_y) = \bar{Q}(t)[1 - F(y)],$$
$$\bar{\mathcal{Z}}(t)(A_y) = \mathcal{Z}(A_y + \bar{S}(t)) + \int_0^t \nu(A_y + \bar{S}(s,t))d\bar{B}(s).$$

This only defines $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ for Borel sets of the form (y, ∞). By the π-λ theorem it defines the measure for all Borel sets in $\mathbb{R}_+$. It is clear by the first equation that $\bar{Q}(t) = \langle 1, \bar{\mathcal{Q}}(t)\rangle$. Plug A_0 in both sides of the second equation in the above to get

$$\begin{aligned}\langle 1, \bar{\mathcal{Z}}(t)\rangle &= \bar{\mathcal{Z}}(A_0) \\ &= \mathcal{Z}(A_{\bar{S}(t)}) + \int_0^t [1 - F(\bar{S}(t) - \bar{S}(s))]d[\lambda s - \bar{Q}(s)] \\ &= \bar{Z}(t),\end{aligned}$$

where the last equality is due to (3.17). So $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$ satisfies the definition of a fluid model solution, implying the existence. The measure $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ will never be zero on $[0, t')$ because of (3.24) and (3.25).

To prove uniqueness, assume there is another solution $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$ for the same initial condition. By Definition 3.1 it must satisfy (3.9)–(3.13). Let

$$t^\dagger = \inf\{t \geq 0 : \bar{X}^\dagger(t) > 0\}.$$

We know that $t^\dagger > 0$ by right-continuity of $\bar{X}^\dagger(t)$ and the non-zero initial condition. So $\bar{S}^\dagger(\cdot)$ has inverse $\bar{T}^\dagger(\cdot)$ on $[0, t^\dagger]$. Let

$$x^\dagger(u) = \bar{X}^\dagger(\bar{T}^\dagger(u)) \text{ for } 0 \leq u \leq \bar{T}^\dagger(t^\dagger).$$

By (3.9)–(3.13), $x^\dagger(\cdot)$ must satisfy (3.23) on $[0, \bar{T}^\dagger(t^\dagger)]$. Due to the uniqueness of solutions to (3.23),

$$x^\dagger(u) = x(u) \quad \text{for } u \leq \min(\bar{T}^\dagger(t^\dagger), a').$$

We first claim that $\bar{T}^\dagger(t^\dagger) \geq a'$. Otherwise $\bar{X}^\dagger(t^\dagger) < a' \leq a$. By (3.24),

$$\bar{X}^\dagger(t^\dagger) = x^\dagger(\bar{T}^\dagger(t^\dagger)) = x(\bar{T}^\dagger(t^\dagger)) > 0,$$

which contradicts the definition of $t^\dagger$. So $x^\dagger(\cdot)$ and $x(\cdot)$ agree on the interval $[0, a']$, which implies that

$$\frac{d}{du}\bar{T}(u) = q(u) \wedge K = q^\dagger(u) \wedge K = \frac{d}{du}\bar{T}^\dagger(u).$$

Since both $\bar{T}(u)$ and $\bar{T}^\dagger(u)$ are absolutely continuous, $\bar{T}^\dagger(u) = \bar{T}(u)$ for all $u \leq a'$. This means that $\bar{X}^\dagger(t) = \bar{X}(t)$ and $\bar{S}^\dagger(t) = \bar{S}(t)$ and for all $t \leq t'$. By (3.9) and (3.10), $(\bar{\mathcal{Q}}^\dagger(t), \bar{\mathcal{Z}}^\dagger(t)) = (\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t))$ for all $t \leq t'$. Uniqueness is proved. □

So far we have established the existence, uniqueness of fluid model solution on a non-trivial interval $[0, t']$. The following "restarting" lemma helps to extend the result in Lemma 3.1 to a larger interval.

Lemma 3.2. *Assume* (3.15) *and* (3.16). *Let* $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot))$ *be a solution to the fluid model* (K, λ, ν) *on the interval* $[0, t_1]$ *for some* $t_1 > 0$. *If* $(\bar{\mathcal{Q}}_2(\cdot), \bar{\mathcal{Z}}_2(\cdot))$ *is a solution to the fluid model with initial condition* $(\bar{\mathcal{Q}}_1(t_1), \bar{\mathcal{Z}}_1(t_1))$ *on the interval* $[0, t_2]$ *for some* $t_2 > 0$, *then* $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *is a fluid model solution on* $[0, t_1 + t_2]$, *where*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = \begin{cases} (\bar{\mathcal{Q}}_1(t), \bar{\mathcal{Z}}_1(t)) & \text{if } t \in [0, t_1], \\ (\bar{\mathcal{Q}}_2(t_1 + t), \bar{\mathcal{Z}}_2(t_1 + t)) & \text{if } t \in [t_1, t_1 + t_2]. \end{cases}$$

Proof. The proof of this lemma is very straightforward. It is clear that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies the fluid dynamic equations on the interval $[0, t_1]$. For any $t \in (t_1, t_1 + t_2]$, plugging t_1 and $t_1 + (t - t_1)$ in to (3.9) and (3.10) and then taking the summation gives

$$\begin{aligned} \bar{\mathcal{Q}}(t)(A_y) &= \bar{\mathcal{Q}}(0)(A_y) + [\bar{Q}(t) - \bar{Q}(0)]\nu(A_y), \\ \bar{\mathcal{Z}}(t)(A_y) &= \bar{\mathcal{Z}}(0)(A_y + \bar{S}(t) - \bar{S}(0)) \\ &\quad + \int_0^t \nu(A_y + \bar{S}(t) - \bar{S}(s))d[\lambda s - \bar{Q}(s)], \end{aligned}$$

for all $A_y = (y, \infty)$, $y \geq 0$. So $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies the fluid dynamic equations on the interval $[0, t_1 + t_2]$. Clearly, it also satisfies all the constraints (3.11)–(3.13). $\square$

Lemma 3.3. *Assume* (3.15) *and* (3.16). *There exists a unique solution* $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *to the fluid model* (K, λ, ν) *satisfying the non-zero initial condition* $(\xi, \mu) \in \mathscr{I}$ *on the interval* $[0, t^*)$, *where either* $t^* < \infty$ *or* $t^* = \infty$; *in the case when* $t^* < \infty$, *the existence and uniqueness can be extended to* $[0, t^*]$ *with* $(\bar{\mathcal{Q}}(t^*), \bar{\mathcal{Z}}(t^*)) = (\mathbf{0}, \mathbf{0})$. *In both cases,*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) \neq (\mathbf{0}, \mathbf{0}) \quad \textit{for all } t \in [0, t^*).$$

Proof. Lemma 3.1 establishes the existence and uniqueness on a small interval $[0, t_1']$, where

$$t_1' = \bar{T}(a_1'), \tag{3.26}$$

$$a_1' = \sup\{u \leq b : x(u) > 0\}, \tag{3.27}$$

according to (3.24) and (3.25) in the proof of Lemma 3.1, and the constant b is the same as in Lemma A.1 and only depends on ρ and F. We put the subscript 1 on the quantities corresponding to the first piece. Lemma 3.1 also says that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \neq (\mathbf{0}, \mathbf{0})$ on the interval $[0, t_1')$. If $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1')) = (\mathbf{0}, \mathbf{0})$, then let $t^* = t_1'$ and the proof is done and we stop. If $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1')) \neq (\mathbf{0}, \mathbf{0})$, then by (3.27),

$$a_1' = b. \tag{3.28}$$

Viewing $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1'))$ as an initial condition, by Lemma 3.1, There exists a unique fluid model solution $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot))$ on the interval $[0, t_2']$, and similar to (3.26) and (3.27),

$$t_2' = \bar{T}_2(a_2'),$$

$$a_2' = \sup\{u \leq b : x_1(u) > 0\},$$

where $\bar{T}_1(\cdot)$ is the corresponding time change based on $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot))$ (defined in the same way as $\bar{T}(\cdot)$ for the process $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$) and $x_1(\cdot)$ is the solution to (3.23) with $h(\cdot)$ generated by the initial condition $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1'))$ via (3.21). Again, according to Lemma 3.1, $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot)) \neq (\mathbf{0}, \mathbf{0})$ on the interval $[0, t_2')$. By Lemma 3.2, we obtain a fluid model solution on the interval $[0, t_1' + t_2']$ by defining $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = ((\bar{\mathcal{Q}}_1(t - t_1'), \bar{\mathcal{Z}}_1(t - t_1')))$ for all $t \in (t_1', t_1' + t_2']$. If $(\bar{\mathcal{Q}}(t_1' + t_2'), \bar{\mathcal{Z}}(t_1' + t_2')) = (\mathbf{0}, \mathbf{0})$, then let $t^* = t_1' + t_2'$ and the proof is complete. Otherwise we have

$$a_2' = b$$

and we can continue the procedure.

If this procedure never stops, then we get a sequence $\{t_i', a_i'\}_{i=1}^{\infty}$ with $a_i' = b$ for all i. Setting

$$t^* = \sum_{i=1}^{\infty} t_i',$$

we have established the existence of a fluid model solution on the interval $[0, t^*)$; the solution never reaches zero before t^*. If $\sum_{i=1}^{\infty} t_i' = \infty$, the proof is complete because the whole interval $[0, \infty)$ is covered. Otherwise, for each $0 \leq s < t^*$, there exists an i_s such that $\sum_{i=i_s}^{\infty} t_i' \geq s$. So

$$\lim_{t \to t^*} \bar{S}(s, t) > \sum_{i=i_s}^{\infty} a_i' = \sum_{i=i_s}^{\infty} b = \infty.$$

By the fluid dynamic equation (3.10), $\lim_{t \to t^*} \bar{\mathcal{Z}}(t) = \mathbf{0}$. The constraints (3.12) and (3.13) implies $\lim_{t \to t^*} \bar{\mathcal{Q}}(t) = \mathbf{0}$. So we can extend the existence to of the fluid model solution to the interval $[0, t^*]$ with $(\bar{\mathcal{Z}}(t^*), \bar{\mathcal{Z}}(t^*)) = (\mathbf{0}, \mathbf{0})$. We have now established the existence of fluid model solution. To prove the uniqueness, note that the interval $[0, t^*)$ is covered by $\bigcup_{j=0}^{\infty} [\sum_{i=0}^{j} t_i', \sum_{i=0}^{j+1} t_i']$ (here we take $t_0 = 0$ for notational convenience). The uniqueness of the solution on the interval $[0, t_1']$ follows directly from Lemma 3.1. The uniqueness on the interval $[t_1', t_1' + t_2']$ can be proved using the same argument in Lemma 3.1 by viewing $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1')$ as the

initial condition and $(\bar{\mathcal{Q}}(t_1' + \cdot), \bar{\mathcal{Z}}(t_1' + \cdot)$ as the corresponding fluid model solution on the interval $[0, t_2']$. Continuing with this procedure establish the uniqueness. This completes the proof. □

The following lemma establishes the workload conserving property for any fluid model solution before it reaches zero.

Lemma 3.4. *Assume* (3.15) *and* (3.16). *For the fluid model solution in Lemma 3.3, we have the following* workload conserving *property on* $[0, t^*)$*:*

$$\langle \chi, \bar{\mathcal{Q}}(t) \rangle + \langle \chi, \bar{\mathcal{Z}}(t) \rangle = \langle \chi, \xi \rangle + \langle \chi, \mu \rangle + (\rho - 1)t. \tag{3.29}$$

Proof. By (3.9) and (3.10), we have

$$\begin{aligned} \langle \chi, \bar{\mathcal{Q}}(t) \rangle &= \int_0^\infty \bar{\mathcal{Q}}(t)(A_y)dy = \bar{Q}(t)\beta, \\ \langle \chi, \bar{\mathcal{Z}}(t) \rangle &= \int_0^\infty \mu(A_y + \bar{S}(t))dy \\ &\quad + \int_0^\infty \int_0^t \nu(A_y + \bar{S}(s,t))d[\lambda s - \bar{Q}(s)]dy. \end{aligned} \tag{3.30}$$

Let $\tilde{F}$ be the distribution function associated with the probability measure $\frac{1}{\langle 1, \mu \rangle}\mu$, so that $\mu(A_y) = \bar{Z}(0)[1 - F(y)]$. Since the cumulative service amount $\bar{S}(\cdot)$ has an inverse on interval $[0, t^*)$, we can perform the change of variable $u = \bar{S}(t)$ and $t = \bar{T}(u)$ for all $t < t^*$. The first term in (3.30) becomes

$$\begin{aligned} &\bar{Z}(0) \int_0^\infty [1 - F(y)]dy + \bar{Z}(0) \int_0^\infty [F(y) - F(y+u)]dy \\ &= \langle \chi, \mu \rangle + \bar{Z}(0) \int_0^u -[1 - F(v)]dv \\ &= \langle \chi, \mu \rangle - \int_0^u \mathcal{Z}(A_v)dv, \end{aligned} \tag{3.31}$$

and the second term in (3.30), by applying Fubini's theorem, becomes

$$\begin{aligned}
&\int_0^u \int_0^\infty \nu(A_y + u - v)dyd[\lambda\bar{T}(v) - \bar{Q}(T(v))] \\
&= \beta \int_0^u \int_0^\infty \frac{1 - F(y + u - v)}{\beta} dyd[\lambda\bar{T}(v) - \bar{Q}(T(v))] \\
&= \beta \int_0^u [1 - F_e(u - v)]d[\lambda\bar{T}(v) - \bar{Q}(T(v))] \\
&= \lambda\beta\bar{T}(u) - \beta[\bar{Q}(T(u)) - \bar{Q}(0)] \\
&\quad - \beta \int_0^u F_e(u - v)d[\lambda\bar{T}(v) - \bar{Q}(T(v))].
\end{aligned} \tag{3.32}$$

To deal with the last term in the above, perform the change of variable $u = \bar{S}(t)$ and $t = \bar{T}(u)$ for (3.17). Note that $\bar{T}'(u) = \bar{Z}(\bar{T}(u)) = \bar{Z}(t)$, So we have

$$\begin{aligned}
\bar{T}'(u) &= \mu(A_u) + \beta \int_0^u \frac{1 - F(u - v)}{\beta} d[\lambda\bar{T}(v) - \bar{Q}_B(\bar{T}(v))] \\
&= \mu(A_u) - \beta \int_0^u F_e'(u - v)d[\lambda\bar{T}(v) - \bar{Q}_B(\bar{T}(v))].
\end{aligned}$$

Integrating both sides of the above equation yields

$$\bar{T}(u) = \int_0^u \mu(A_v)dv - \beta \int_0^u F_e(u - v)d[\lambda\bar{T}(v) - \bar{Q}_B(\bar{T}(v))]. \tag{3.33}$$

The proof is completed by combining (3.31), (3.32) and (3.33) and substituting $\bar{T}(u)$ with t. □

3.2.2 Starting with Zero Initial Condition when $\rho \leq 1$

Intuitively, the fluid model solution should stay at zero for ever in this case. We rigorously prove this result in the following lemma.

Lemma 3.5. *When $\rho \leq 1$, $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv (\mathbf{0}, \mathbf{0})$ is the unique solution to the fluid model (λ, K, ν) with initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$.*

Proof. Since $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \mathbf{0}$, $\bar{Z}(\cdot) \equiv 0$. By (3.8) we have $\bar{S}(s,t) = \infty$ for all $t > s \geq 0$. So $\nu(A_y + \bar{S}(s,t)) = 0$ for all $x \geq 0$ since ν can not have any mass at infinity. This implies that the integral on the right hand side of (3.10) is zero.

So $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \mathbf{0}$ satisfies equation (3.10). It is clear that fluid dynamic equation (3.9) and constraints (3.11) through (3.13) are satisfied. So $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \mathbf{0}$ is a fluid model solution.

We now prove that it is the only solution. If $(\mathbf{0}, \mathbf{0})$ is the unique fluid model solution on the interval $[0, K/\lambda]$, then by Lemma 3.2 we can extend the uniqueness to $[K/\lambda, 2K/\lambda]$ and so on to $[0, \infty)$. Otherwise, there is another solution on $[0, K/\lambda]$ which is denoted by $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$. By (3.9) and (3.10), for any fluid model solution $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$ starting at $(\mathbf{0}, \mathbf{0})$, we have

$$\begin{aligned}\bar{X}^\dagger(t) &= \bar{\mathcal{Q}}^\dagger([0,\infty) + \bar{\mathcal{Z}}^\dagger([0,\infty) \\ &\leq Q^\dagger(t)\beta + \int_0^t 1d[\lambda s - \bar{Q}^\dagger(s)] \leq \lambda t.\end{aligned}$$

So $\bar{Q}^\dagger(\cdot) \equiv 0$ on the interval $[0, K/\lambda]$ by (3.12). Thus, by (3.10), the workload process

$$\begin{aligned}\bar{W}^\dagger(t) &= \int_0^\infty \int_0^t \nu(A_y + \bar{S}^\dagger(s,t))d\lambda s dy \\ &= \lambda \int_0^t \int_0^\infty \nu(A_y + \bar{S}^\dagger(s,t))dy ds\end{aligned}$$

for all $t \in [0, K/\lambda]$, where the second inequality is due to Fubini's theorem. Since

$$\int_0^\infty \nu(A_y + \bar{S}^\dagger(s,t))dy \leq \int_0^\infty \nu(A_y)dy < \infty,$$

$\bar{W}^\dagger(\cdot)$ is continuous on $[0, K/\lambda]$. This is a solution starting from zero, so

$$\bar{W}^\dagger(0) = 0. \tag{3.34}$$

But it is different from $(\mathbf{0}, \mathbf{0})$, so there must be a $t_1 \in (0, K/\lambda]$ such that $(\bar{\mathcal{Q}}^\dagger(t_1), \bar{\mathcal{Z}}^\dagger(t_1)) \neq (\mathbf{0}, \mathbf{0})$, which implies that

$$\bar{W}^\dagger(t_1) > 0. \tag{3.35}$$

Let $t_0 = \sup_{0 \leq t < t_1}\{\bar{W}^\dagger(t) > 0\}$, then $0 \leq t_0 < t_1$ by (3.34) and (3.35) and continuity of $\bar{W}^\dagger(\cdot)$. Again by continuity of $\bar{W}^\dagger(\cdot)$, there exists a $t_\delta \in (t_0, t_1)$ such that $\bar{W}^\dagger(t_\delta) =$

$\delta < \bar{W}^\dagger(t_1)$ for some $\delta > 0$. On the interval $[t_\delta, t_1]$, $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$ never reaches zero. So by Lemma 3.2 and 3.4,

$$\bar{W}^\dagger(t_\delta + t) = \bar{W}^\dagger(t_\delta) + (1-\rho)t \quad \text{for } t \in [0, t_1 - t_\delta].$$

This implies that $\bar{W}^\dagger(t_1) \leq \bar{W}^\dagger(t_\delta)$, which is a contradiction. □

3.2.3 Starting with Zero Initial Condition when $\rho > 1$

In Jean-Marie and Robert [31] and Puha, Stolyar and Williams [44], a very nice approach has been developed for overloaded PS queue with zero initial condition. We can apply the same approach to the LPS queue without much adjustment, since the fluid models of the LPS queue and PS queue behave the same until the time that total job size becomes larger than K.

Intuitively, the fluid model solution should grow as time goes by. Let us first assume that the fluid queue length process $\bar{X}(\cdot)$ grows linearly on a small interval, i.e.

$$\begin{aligned} \bar{Q}(t) &= \langle 1, \bar{\mathcal{Q}}(t) \rangle = 0, \\ \bar{Z}(t) &= \langle 1, \bar{\mathcal{Z}}(t) \rangle = mt, \end{aligned} \tag{3.36}$$

for all $t \in [0, K/m]$, where $m > 0$ is to be determined. The following analysis is taken from [44]. By (3.13)

$$\bar{S}(s,t) = \int_s^t \frac{1}{m\tau} d\tau = \frac{1}{m} \log \frac{t}{s}, \quad 0 < s < t \leq K/m. \tag{3.37}$$

Plug (3.36) and (3.37) in equation (3.17) to get

$$mt = \lambda \int_0^t [1 - F(\frac{1}{m} \log \frac{t}{s})] ds \quad \text{for all } t \leq \frac{K}{m}.$$

Perform the change of variable $v = \frac{1}{m} \log \frac{t}{s}$ to obtain

$$\frac{1}{\lambda\beta} mt = mt \int_0^\infty e^{-mv} \frac{1 - F(v)}{\beta} dv \quad \text{for all } t \leq \frac{K}{m}.$$

By the definition of F_e and ρ, we must have

$$\int_0^\infty e^{-mv} dF_e(v) = \frac{1}{\rho}. \tag{3.38}$$

Note that the left hand side is the Laplace transform of the distribution F_e. As a function of $m \in (0, \infty)$, it is strictly decreasing and maps onto $(0, 1)$. Since $\rho > 1$ in this case, (3.38) has a unique solution, which we denote by m_ρ^*. Now let

$$\begin{aligned} &\bar{\mathcal{Q}}(t)(A_y) = 0, \\ &\bar{\mathcal{Z}}(t)(A_y) = \lambda \int_0^t [1 - F(y + \frac{1}{m_\rho^*} \log \frac{t}{s})] ds \end{aligned} \tag{3.39}$$

for all $t \in [0, \frac{K}{m_\rho^*}]$ and $y \geq 0$. It is clear that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is a fluid model solution on the interval $[0, \frac{K}{m_\rho^*}]$. By Lemma 3.2, $(\bar{\mathcal{Q}}(\frac{K}{m_\rho^*} + \cdot), \bar{\mathcal{Z}}(\frac{K}{m_\rho^*} + \cdot))$ can be viewed as the fluid model solution with initial condition $(\bar{\mathcal{Q}}(\frac{K}{m_\rho^*}), \bar{\mathcal{Z}}(\frac{K}{m_\rho^*}))$, which exists on $[0, \infty)$. Thus, we have found a fluid model solution with zero initial condition.

Similar as in the case $\rho \leq 1$, the difficulty is to prove uniqueness. [44] has established existence and uniqueness of fluid model solutions for overloaded PS queue with zero initial condition. We can borrow the result for the reason that the total fluid amount of jobs of any fluid model solution starting at zero is bounded by λt at any time $t \geq 0$, as explained in the proof of Lemma 3.5. The sharing limit K is never reached on the interval $[0, K/\lambda]$, so the model is the same as a standard PS queue. In fact, the fluid dynamic equation (3.10) is what is need in Theorem 4.2 and Lemma 4.10, which implies uniqueness on the interval $[0, K/\lambda]$. The uniqueness can be extended to $[K/\lambda, \infty)$ by Lemma 3.2, since $\bar{X}(K/\lambda) > 0$. So we have the following result.

Lemma 3.6. *Assume* (3.15) *and* (3.16). *When* $\rho > 1$, *there exists a unique solution to the fluid model* (K, λ, ν) *with initial condition* $(\mathbf{0}, \mathbf{0})$.

We are now in a position to sum up all the above cases and prove all results on the fluid model.

Proof of Theorem 3.1. If the initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$, then the result is established by Lemma 3.5 and Lemma 3.6. If $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$, then by Lemma 3.3 either we have existence and uniqueness on the interval $[0, \infty)$ and the proof is done, or the result holds on a finite interval $[0, t^*]$ with $(\bar{\mathcal{Q}}(t^*), \bar{\mathcal{Z}}(t^*)) = (\mathbf{0}, \mathbf{0})$. The result is then established by applying Lemma 3.2 and Lemma 3.5. □

Proof of Proposition 3.1. If $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$ and $\rho \leq 1$, then it follows from Lemma 3.5 that $\bar{W}(t) = (0 + (\rho - 1))^+$. If $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$ and $\rho > 1$, for any $t \in [0, K/m^*_\rho]$, take the integration of both sides of (3.39) with respect to y to get

$$\bar{W}(t) = \langle \chi, \bar{\mathcal{Z}}(t) \rangle = \lambda \int_0^\infty \int_0^t [1 - F(y + \frac{1}{m^*_\rho} \log \frac{t}{s})] ds dy$$

Performing the change of variable $v = \frac{1}{m^*_\rho} \log \frac{t}{s}$ and applying Fubini's theorem, we obtain $\bar{W}(t) = 0 + (\rho - 1)t$. If $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$, then the workload conserving property holds before the fluid model solution reaches zero. Note that the fluid model solution reaches zero if and only if the workload reaches zero. So when $\rho \geq 1$, $\bar{W}(t) = w + (\rho - 1)t > 0$ for all $t > 0$ and the result holds on $[0, \infty)$, When $\rho < 1$, $\bar{W}(t_w) = 0$ for $t_w = w/(1 - \rho)$. By weak stability, $\bar{W}(t) = 0$ for all $t \geq t_w$. □

Proof of Theorem 3.2. Weak stability is already proved in Lemma 3.5. Since the descriptor $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t))$ equals $(\mathbf{0}, \mathbf{0})$ if and only if $\bar{W}(t) = 0$, the stability follows immediately from Proposition 3.1. □

3.3 Convergence to Equilibrium States for Fluid Model

By the workload conservation property (Proposition 3.1), starting from a valid initial state with $w = \langle \chi, \xi + \mu \rangle < \infty$, the workload of fluid model solution will blow up to infinity if $\rho > 1$; and the fluid model solution will be zero after a finite time if $\rho < 1$. So the only interesting case to study long term behavior will be the case where the queue is critically loaded. Now let consider the case where the traffic intensity

$$\rho = \lambda \beta = 1. \tag{3.40}$$

In other words, our fluid model is critically loaded.

Definition 3.3. *An element* $(\xi, \mu) \in \mathscr{I}$ *is called an* equilibrium state *for the fluid model* (K, λ, ν) *if the solution to the fluid model with initial condition* (ξ, μ) *satisfies*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = (\xi, \mu) \quad \textit{for all } t \geq 0.$$

Denote

$$\beta_e = \langle \chi, \nu_e \rangle,$$

where ν_e is the *equilibrium* measure of ν, i.e. $\nu_e([0, x]) = \frac{1}{\beta} \int_0^x \nu((y, \infty))dy$ for all $x \geq 0$. We have the following definition.

Definition 3.4. *Let* $\Delta_{K,\nu} : \mathbb{R}_+ \to \mathbf{M}_1 \times \mathbf{M}_2$ *be the lifting map associated with the probability measure* ν *and constant* K *given by*

$$\Delta_{K,\nu} w = \Big(\frac{(w - K\beta_e)^+}{\beta} \nu, \frac{w \wedge K\beta_e}{\beta_e} \nu_e \Big) \quad \textit{for } w \in \mathbb{R}_+.$$

The main objective of this section is to show the following long-term behavior of the critically loaded fluid model, which helps to establish the state space collapse in Chapter 5.

Theorem 3.3. *Assume* (3.15)–(3.40) *and*

$$\nu \textit{ is non-lattice.} \tag{3.41}$$

The unique solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *to the fluid model* (K, λ, ν) *with a valid initial state* (ξ, μ) *such that* $w = \langle \chi, \xi + \mu \rangle < \infty$ *satisfies*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) \to \Delta_{K,\nu} w \quad \textit{as } t \to \infty.$$

Moreover, for fixed constants $p, M > 0$ *the convergence is uniform for all initial conditions in the set*

$$\mathscr{I}_M^p = \{\xi \in \mathscr{I} : \langle \chi, \xi + \mu \rangle < M, \quad \langle \chi^{1+p}, \xi + \mu \rangle < M\}. \tag{3.42}$$

Section 3.3.1 characterizes the equilibrium states for the fluid model. Section 3.3.2 presents the proof of convergence (the first half of Theorem 3.3), and Section 3.3.3 presents the proof of uniform convergence (the second half of Theorem 3.3).

3.3.1 Equilibrium States

Our first result is a characterization of an equilibrium state.

Theorem 3.4. *An element $(\xi, \mu) \in \mathscr{I}$ is an equilibrium state if and only if*

$$(\xi, \mu) = \Delta_{K,\nu} w \quad \textit{for some } w \in [0, \infty). \tag{3.43}$$

Proof. Suppose $(\xi, \mu) = \Delta_{K,\nu} w$ for some $w \in [0, \infty)$, we need show that

$$(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \Delta_{k,\nu} w = (\frac{(w - K\beta_e)^+}{\beta}\nu, \frac{w \wedge K\beta_e}{\beta_e}\nu_e)$$

is the fluid model solution. If $w = 0$, then by weak stability, $\Delta_{k,\nu} 0 = (\mathbf{0}, \mathbf{0})$ is the fluid model solution. So let us now focus on the case where $w > 0$. The fluid buffer size and queue size are

$$\bar{Q}(t) = \langle 1, \bar{\mathcal{Q}}(t) \rangle = \frac{(w - K\beta_e)^+}{\beta},$$
$$\bar{Z}(t) = \langle 1, \bar{\mathcal{Z}}(t) \rangle = \frac{w \wedge K\beta_e}{\beta_e}.$$

If $\bar{Z}(t) < K$, then $w < K\beta_e$ which implies that $\bar{Q}(t) = 0$; if $\bar{Q}(t) > 0$, then $w > K\beta_e$ which implies that $\bar{Z}(t) = K$. So condition (3.12) and (3.13) in Definition 3.1 are satisfied. Since $\bar{Q}(t)$ and $\bar{Z}(t)$ remain to be a constant, (3.11) holds trivially. This also implies that the fluid dynamic equation (3.9) is satisfied. It remains to verify the fluid dynamic equation (3.10). The fluid accumulative service amount

$$\bar{S}(t) = \frac{t}{\bar{Z}(0)} = \frac{\beta_e}{w \wedge K\beta_e} t$$

since $\bar{Z}(t)$ is a constant. The right hand side of (3.10) becomes

$$\frac{w \wedge K\beta_e}{\beta_e}\nu_e(A_y + \frac{\beta_e}{w \wedge K\beta_e}t) + \lambda \int_0^t \nu(A_y + \frac{\beta_e}{w \wedge K\beta_e}(t - s))ds,$$

which equals $\frac{w\wedge K\beta_e}{\beta_e}\nu_e(A_y) = \bar{\mathcal{Z}}(t)(A_y)$, for all $x \geq 0$. So (3.10) is verified and $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is the fluid model solution.

Suppose that (ξ, μ) is an equilibrium state, we need to show that (ξ, μ) takes the form (3.43). If $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$, then trivially $(\xi, \mu) = \Delta_{K,\nu}0$. Let us now assume that $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$. Since $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot) \equiv (\xi, \mu)$ is the fluid model solution, the fluid dynamic equation (3.10) must be satisfied. i.e.

$$\mu(A_y) = \mu(A_y + \frac{t}{\langle 1, \mu\rangle}) + \lambda \int_0^t \nu(A_y + \frac{t-s}{\langle 1, \mu\rangle})ds,$$

for all $x, t \geq 0$. Differentiation with respect to t shows that $\mu = \langle 1, \mu\rangle\nu_e$. Since (ξ, μ) is a valid state, $\xi = \langle 1, \xi\rangle\nu$. Let

$$w = \langle \chi, \xi + \mu\rangle = \langle 1, \mu\rangle\beta + \langle 1, \xi\rangle\beta_e.$$

Again by validity of state (ξ, μ), $\langle 1, \xi\rangle = \frac{(w-K\beta_e)^+}{\beta}$ and $\langle 1, \mu\rangle = \frac{w\wedge K\beta_e}{\beta_e}$. So we conclude that $(\xi, \mu) = \Delta_{K,\nu}w$. □

3.3.2 Convergence to Equilibrium States

We now identify conditions under which the fluid model solution starting at a valid initial condition (ξ, μ) will converge to an equilibrium state.

By the fluid dynamic equation (3.9), $\bar{Q}(t)\beta = \langle \chi, \bar{\mathcal{Q}}(t)\rangle \leq w$. It follows from the workload conservation property that $\bar{W}(t) \equiv w = \langle \chi, \xi + \mu\rangle \geq \langle \chi, \bar{\mathcal{Q}}(t)\rangle$ for all $t \geq 0$. So

$$\bar{Q}(t) = (\bar{X}(t) - K)^+ \leq \frac{w}{\beta} \quad \text{for all } t \geq 0. \tag{3.44}$$

Since $\bar{W}(t) = 0$ if and only if $\bar{Z}(t) = 0$,

$$\bar{Z}(t) = (\bar{X}(t) \wedge K) > 0 \quad \text{for all } t \geq 0. \tag{3.45}$$

So the function $\bar{S}(\cdot)$ as defined in (3.6) has an inverse on the interval $[0, \infty)$, which is denoted by $\bar{T}(\cdot)$.

If the initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$, then by weak stability, the fluid model solution will always be zero. So $(\mathbf{0}, \mathbf{0})$ is an equilibrium state. From now on, we focus on the case where the initial condition $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$. For the valid initial condition (ξ, μ) and an $\epsilon \in (-1, 1)$, define the ϵ-perturbation of it by

$$(\xi_\epsilon, \mu_\epsilon) = \begin{cases} \Big((1 + \frac{K - \langle 1, \xi \rangle}{\langle 1, \xi \rangle} \wedge \epsilon)\xi, \mu\Big) & \text{if } \langle 1, \xi \rangle < K, \\ \Big(\xi, (1 + \epsilon)\mu\Big) & \text{if } \mu \neq \mathbf{0}, \\ \Big((1 + (\epsilon \wedge 0))\xi, \mu + (\epsilon - 0)^+ \xi\Big) & \text{if } \langle 1, \xi \rangle = K, \mu = \mathbf{0}. \end{cases}$$

It is clear that $(\xi_\epsilon, \mu_\epsilon)$ is still a valid initial condition. Let $x^\epsilon(\cdot)$ denote the solution to (3.23) with $h_{\xi,\mu}$ replaced by $h_{\xi_\epsilon, \mu_\epsilon}$. We have the following comparison.

Lemma 3.7. *Assume* (3.15)–(3.16). *For all* $\epsilon \in (0, 1)$,

$$x^{-\epsilon}(u) < x(u) < x^\epsilon(u) \quad \textit{for all } u \geq 0.$$

Proof. Let $u^* = \inf\{u \geq 0 : x(u) > x^\epsilon(u)\}$. To prove $x(u) < x^\epsilon(u)$, it is enough to show that $u^* = \infty$. Note that $x(0) = h_{\xi,\mu}(0) < h_{\xi_\epsilon, \mu_\epsilon}(0) = x^\epsilon(0)$. By right-continuity of $x(\cdot)$ and $x^\epsilon(\cdot)$, $u^* > 0$. Suppose $u^* < \infty$ and consider the following difference at u^*,

$$\begin{aligned} & x^\epsilon(u^*) - x(u^*) \\ \geq & \int_0^{u^*} [(x^\epsilon(u^* - v) - K)^+ - (x(u^* - v) - K)^+] dF(v) \\ & + \int_0^{u^*} [(x^\epsilon(u^* - v) \wedge K) - (x(u^* - v) \wedge K)] dF_e(v). \end{aligned} \tag{3.46}$$

Assumption (3.16) implies that $F(0) < 1$. So there exists $u' \in (0, u^*)$ such that

$$\int_{u'-\delta}^{u'+\delta'} dF(v) > 0, \quad \int_{u'-\delta}^{u'+\delta'} dF_e(v) > 0$$

for all $\delta > 0$. Choose δ small enough such that $0 < u_F - \delta, u_F + \delta < u^*$. By the definition of u^*, we have that

$$\kappa = x^\epsilon(u^* - u_F - \delta) - x(u^* - u_F - \delta) > 0.$$

By right continuity of $x(\cdot)$ and $x^\epsilon(\cdot)$, we can choose δ small enough such that

$$x^\epsilon(u) - x(u) \geq \frac{\kappa}{2} \quad \text{for all } u \in [u^* - u_F - \delta, u^* - u_F + \delta].$$

So by (3.46), we have

$$x^\epsilon(u^*) - x(u^*) \geq \frac{\kappa}{2} \min\left(\int_{u'}^{u'+\delta'} dF(v), \int_{u'}^{u'+\delta'} dF_e(v)\right) > 0.$$

This contradicts the definition of u^*. So we must have that $u^* = \infty$. The proof for the other inequality is completely analogous. □

For the solution $x(\cdot)$ to (3.23) with initial condition (ξ, μ), define

$$x(\infty) = \frac{1}{\beta}(w - K\beta_e)^+ + \frac{1}{\beta_e}(w \wedge K\beta_e), \tag{3.47}$$

where $w = \langle \chi, \xi + \mu \rangle$. We now use the above lemma and the key renewal theorem to show the following convergence.

Lemma 3.8. *Assume* (3.15)-(3.41) *and* (3.40). *The solution* $x(\cdot)$ *to* (3.23) *with initial condition* (ξ, μ) *satisfies*

$$x(u) \to x(\infty) \quad \textit{as } u \to \infty.$$

Proof. We first study the case where $w = \langle \chi, \xi + \mu \rangle > K\beta_e$. Convolve both sides of (3.23) with $U(\cdot)$, the renewal function of $F(\cdot)$, to get

$$x * U(u) = h_{\xi,\mu} * U(u) + (x - K)^+ * F * U(u) + (x \wedge K) * F_e * U(u).$$

Since $x = (x - K)^+ + (x \wedge K)$, by moving all terms containing $(x - K)^+$ to the left and all terms containing $(x \wedge K)$ to the right, we obtain

$$\begin{aligned}(x(u) - K)^+ = h_{\xi,\mu} * U(u) - K(1 - F_e) * U(u) \\ + [K - (x \wedge K)] * (1 - F_e) * U(u).\end{aligned} \tag{3.48}$$

Both $h_{\xi,\mu}(\cdot)$ and $1-F_e(\cdot)$ are directly Riemann integrable since they are non-increasing and integrable functions. By the key renewal theorem, we have the convergence of the first two terms on the right hand side of (3.48):

$$\lim_{u\to\infty} h_{\xi,\mu} * U(u) = \frac{w}{\beta},$$
$$\lim_{u\to\infty} K(1-F_e) * U(u) = \frac{K\beta_e}{\beta}.$$

Note that $\frac{w}{\beta} - \frac{K\beta_e}{\beta} > 0$ in this case, and the last term in (3.48) is always non-negative. So there exists $u_1 > 0$ such that

$$(x(u) - K)^+ > 0 \quad \text{for all } u \geq u_1.$$

Equivalently, this means that $K - (x(u) \wedge K) = 0$ for all $u \geq u_1$. So the last term in (3.48) can be bounded by

$$\int_{u-u_1}^{u} K d[(1-F_e) * U(v)] = K[(1-F_e) * U(u) - (1-F_e) * U(u-u_1)]$$

which converges to 0 by the key renewal theorem. So in this case we have $\lim_{t\to\infty} x(u) = \frac{(w-K\beta_e)}{\beta} + K = x(\infty)$.

In the case where $w < K\beta_e$, we convolve both sides of (3.23) with $U_e(\cdot)$, the renewal function of $F_e(\cdot)$ to get

$$x * U_e(u) = h_{\xi,\mu} * U_e(u) + (x-K)^+ * F * U_e(u) + (x \wedge K) * F * U_e(u).$$

By moving all terms containing $(x-K)^+$ to the right and all terms containing $(x \wedge K)$ to the left, we obtain

$$(x(u) \wedge K) = h_{\xi,\mu} * U_e(u) - (x-K)^+ * (1-F) * U_e(u). \tag{3.49}$$

Again, by the key renewal theorem, the first term in the above converges:

$$\lim_{u\to\infty} h_{\xi,\mu} * U_e(u) = \frac{w}{\beta_e}.$$

Note that $\frac{w}{\beta_e} - K < 0$ in this case, and the last term in (3.49) is always non-positive. So there exists $u_2 > 0$ such that

$$(x(u) \wedge K) < K \quad \text{for all } u \geq u_2.$$

Equivalently, this means that $(x(u) - K)^+ = 0$ for all $u \geq u_2$. According to the upper bound (3.44), the absolute value of the last term in (3.49) can be bounded by

$$\int_{u-u_2}^{u} \frac{w}{\beta} d[(1-F) * U_e(v)] = \frac{w}{\beta}[(1-F) * U_e(u) - (1-F) * U_e(u-u_2)]$$

which converges to 0 by the key renewal theorem. So in this case we have $\lim_{t\to\infty} x(u) = \frac{w}{\beta_e} = x(\infty)$.

Now it only remains to study the case where $w = K\beta_e$. For any $\epsilon \in (0,1)$, let $(\xi_\epsilon, \mu_\epsilon)$ denote the ϵ-perturbation of the initial condition (ξ, ϵ), introduced before Lemma 3.7. It is clear that $w_\epsilon = \langle \chi, \xi_\epsilon + \mu_\epsilon \rangle = \int_0^\infty h_{\xi_\epsilon, \mu_\epsilon}(u) du$ satisfies

$$0 < w_\epsilon - \beta_e K \leq \epsilon \beta_e K.$$

Following from the discussion of our first case,

$$\lim_{u\to\infty} x^\epsilon(u) = K + \frac{(w_\epsilon - \beta_e K)^+}{\beta}.$$

By Lemma 3.7, $x(u) < x^\epsilon(u)$ for all $u \geq 0$. So for all $\epsilon > 0$ there exists u_1' such that when $u \geq u_1'$

$$x(u) \leq K + \frac{(w_\epsilon - \beta_e K)^+}{\beta} + \epsilon \leq K + (\frac{\beta_e K}{\beta} + 1)\epsilon.$$

Similarly, we introduce the $-\epsilon$-perturbation $(\xi_{-\epsilon}, \mu_{-\epsilon})$. It is clear that $w_{-\epsilon} = \langle \chi, \xi_{-\epsilon} + \mu_{-\epsilon} \rangle = \int_0^\infty h_{\xi_{-\epsilon}, \mu_{-\epsilon}}(u) du$ satisfies

$$-\epsilon \beta_e K < w_{-\epsilon} - \beta_e K < 0.$$

Following from the discussion of our first case,

$$\lim_{u\to\infty} x^{-\epsilon}(u) = \frac{w_{-\epsilon}}{\beta_e}.$$

By Lemma 3.7, $x(u) > x^{-\epsilon}(u)$ for all $u \geq 0$. So for all $\epsilon > 0$ there exists u_2' such that when $u \geq u_2'$

$$x(u) \geq \frac{w_{-\epsilon}}{\beta_e} - \epsilon \geq K - (K+1)\epsilon.$$

Summarizing this case, we have $\lim_{t\to\infty} x(u) = K = x(\infty)$. □

Lemma 3.9. *Assume* (3.15)–(3.41) *and* (3.40). *Let* $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *be the solution to the fluid model* (K, λ, ν) *with initial condition* (ξ, μ). *Let* $w = \langle \chi, \mu \rangle$. *We have as* $t \to \infty$,

$$\mathbf{d}[\bar{\mathcal{Q}}(t), \frac{(w - K\beta_e)^+}{\beta}\nu] \to 0, \tag{3.50}$$

$$\sup_{y\in[0,\infty)} |\bar{\mathcal{Z}}(t)(A_y) - \frac{w \wedge K\beta_e}{\beta_e}\nu_e(A_y)| \to 0. \tag{3.51}$$

Proof. If $w = 0$, the result holds trivially. Now assume that $w \neq 0$. Let

$$q(\infty) = (x(\infty) - K)^+ = \frac{(w - K\beta_e)^+}{\beta},$$
$$z(\infty) = x(\infty) \wedge K = \frac{w \wedge K\beta_e}{\beta_e}.$$

Using the fluid dynamic equation (3.9), for all Borel set $A_y \subset \mathbb{R}_+$ we have

$$|\bar{\mathcal{Q}}(t)(A_y) - q(\infty)\nu(A_y)| \leq |\bar{Q}(t) - q(\infty)|.$$

By the change of variable $u = \bar{S}(t)$ $(t = \bar{T}(u))$ and the definition of the Prohorov metric,

$$\mathbf{d}[\bar{\mathcal{Q}}(t), \frac{(w - K\beta_e)^+}{\beta}\nu] \leq |\bar{q}(u) - q(\infty)|.$$

By Lemma 3.8, there exists $u_1 > 0$ such that when $u > u_1$ we have $|\bar{q}(u) - q(\infty)| < \epsilon$. So for all $\epsilon > 0$, there exists $t_1 = Ku_1 \geq \bar{T}(u_1)$ such that

$$\mathbf{d}[\bar{\mathcal{Q}}(t), \frac{(w - K\beta_e)^+}{\beta}\nu] < \epsilon \quad \text{for all } t > t_1. \tag{3.52}$$

It remains to study the limiting behavior of $\bar{\mathcal{Z}}(\cdot)$. Perform the change of variable $u = \bar{S}(t)$ $(t = \bar{T}(u))$ to the fluid dynamic equation (3.10), we get

$$\bar{\mathcal{Z}}(\bar{T}(u))(A_y) = \bar{\mathcal{Z}}(0)(A_y + u) + \int_0^u \nu(A_y + u - v)d[\lambda\bar{T}(v) - q(v)].$$

Since $z(\infty)\nu_e(A_y) = \lambda \int_0^u \nu(A_y + u - v)z(\infty)dv$ and $d\bar{T}(v) = z(v)dv$, the following difference can be written as

$$\begin{aligned} &|\bar{\mathcal{Z}}(\bar{T}(u))(A_y) - z(\infty)\nu_e(A_y)| \\ &= \mu(A_y + u) + \int_0^u \nu(A_y + u - v)dq(v) \\ &\quad + \lambda \int_0^u \nu(A_y + u - v)[z(\infty) - z(v)]dv. \end{aligned} \tag{3.53}$$

It is clear that the first term on the right hand side of (3.53) vanishes as $u \to \infty$. By convergence of $x(\cdot)$, for all $\epsilon > 0$ there exists a u_1 such that $|x(v) - x(\infty)| < \epsilon$ if $v \geq u_1$. For all $\epsilon > 0$, we can choose $u_2 > 0$ such that $1 - F(u_2) < \epsilon$. When $u > u_1 + u_2$, the second term in (3.53) can be written as

$$\int_0^{u-u_2} \nu(A_y + u - v)dq(v) + \int_{u-u_2}^u \nu(A_y + u - v)dq(v),$$

which is bounded above by

$$\begin{aligned} &\leq \epsilon|q(u - u_2) - q(0)| + 1|q(u) - q(u - u_2)| \\ &\leq \epsilon\frac{w}{\beta} + \epsilon, \end{aligned}$$

where the last inequality is due to (3.44). The last term in (3.53) can be written as

$$\begin{aligned} &\lambda \int_0^{u_1} \nu(A_y + u - v)[z(\infty) - z(v)]dv \\ &+ \lambda \int_{u_1}^u \nu(A_y + u - v)[z(\infty) - z(v)]dv, \end{aligned}$$

which is bounded by

$$\begin{aligned} &\leq \lambda \sup_{0 \leq u \leq u_1} |z(u) - z(\infty)|[1 - F(u - u_1)] + \epsilon\lambda \int_{u_1}^u [1 - F(v)]dv \\ &\leq \lambda K\epsilon + \lambda\beta\epsilon, \end{aligned}$$

where the last inequality is due to the bound $z(u) < K$ for all $u \geq 0$. So for all $\epsilon > 0$ there exists a $t_2 = K(u_1 + u_2) \geq \bar{T}(u_1 + u_2)$ such that

$$\sup_{y \in [0,\infty)} |\bar{\mathcal{Z}}(t)(A_y) - z(\infty)\nu_e(A_y)| < \epsilon \quad \text{for all } t \geq t_2. \tag{3.54}$$

□

Proof of Theorem 3.4, part I. Since the collection of subsets $\{(y,\infty) : y \in [0,\infty)\}$ forms a π-system, Theorem 2.2 in [5] and (3.50) in Lemma 3.9 immediately imply that

$$\mathbf{d}[\bar{\mathcal{Z}}(t), \frac{w \wedge K\beta_e}{\beta_e}\nu_e] \to 0 \quad \text{as } t \to \infty.$$

This and (3.51) implies the convergence result in Theorem 3.4. □

3.3.3 Uniform Convergence to Equilibrium States

The convergence in the previous subsection depends on the initial condition ξ. We now show that the convergence is uniform for all initial condition ξ in the set $\mathscr{I}_M^p$ defined in Theorem 3.3.

To emphasize the dependency on the initial condition we use $\Upsilon(\xi,\mu)$ to denote the solution to equation (3.23) with initial condition (ξ,μ), and $\Xi(\xi,\mu)$ to denote the solution to the fluid model (K,λ,ν) with initial condition (ξ,μ).

Lemma 3.10. *Assume* (3.15)-(3.41) *and* (3.40). *For each* $\epsilon > 0$ *there exists an* $l^* > 0$ *such that when* $u \geq l^*$,

$$\sup_{x(\cdot)\in\Upsilon(\mathscr{I}_M^p)} |x(u) - x(\infty)| < \epsilon.$$

Proof. To prove this lemma, we need to adjust the proof of Lemma 3.8 with the assistance of Lemma B.1.

Let $\mathscr{H}_M = \{h_{\xi,\mu} : (\xi,\mu) \in \mathscr{I}_M^p\}$. By the definition of the set $\mathscr{I}_M^p$ in Theorem 3.3, $\mathscr{H}_M$ is the set of non-increasing functions which are uniformly integrable with integration less than M. For any $\epsilon > 0$, divide the set $\mathscr{I}_M^p$ into three parts,

$$\mathscr{I}_M^p = \mathscr{I}_\epsilon^+ \cup \mathscr{I}_\epsilon^0 \cup \mathscr{I}_\epsilon^-,$$

where

$$\mathscr{I}_\epsilon^+ = \{(\xi,\mu) \in \mathscr{I}_M^p : \langle \chi, \xi+\mu\rangle \geq K\beta_e(1+\epsilon)\},$$
$$\mathscr{I}_\epsilon^- = \{(\xi,\mu) \in \mathscr{I}_M^p : \langle \chi, \xi+\mu\rangle \leq K\beta_e(1-\epsilon)\},$$

and $\mathscr{I}_\epsilon^0 = \mathscr{I}_M^p \setminus (\mathscr{I}_\epsilon^+ \cup \mathscr{I}_\epsilon^-)$.

We first focus on the set $\mathscr{I}_\epsilon^+$. By doing the same algebra as in the proof of Lemma 3.8, we see that (3.48) holds for any $(\xi, \mu) \in \mathscr{I}_\epsilon^+$. By Lemma B.1 and the key renewal theorem, there exists a u_1^* such that

$$\sup_{(\xi,\mu)\in\mathscr{I}_\epsilon^+} |h_{\xi,\mu} * U(u) - \frac{\langle \chi, \xi+\mu\rangle}{\beta}| < \frac{K\beta_e}{4\beta}\epsilon,$$
$$|K(1-F_e) * U(u) - K\frac{\beta_e}{\beta}| < \frac{K\beta_e}{4\beta}\epsilon,$$

for all $u \geq u_1^*$. So for the first two terms on the right hand side of (3.48) we have

$$h_{\xi,\mu} * U(u) - K(1-F_e) * U(u) \geq \frac{K\beta_e(1+\epsilon)}{\beta} - K\frac{\beta_e}{\beta} - \frac{K\beta_e}{2\beta}\epsilon > 0,$$

for all $(\xi, \mu) \in \mathscr{I}_\epsilon^+$ and $u > u_1^*$. Note that the last term in (3.48) is always non-negative. So when $u \geq u_1^*$ we have $(x(u)-K)^+ > 0$, (or equivalently $K-(x(u)\wedge K) = 0$) for all $(\xi, \mu) \in \mathscr{I}_\epsilon^+$. So the last term in (3.48) can be bounded by

$$\int_{u-u_1^*}^{u} Kd[(1-F_e) * U(v)] = K[(1-F_e) * U(u) - (1-F_e) * U(u-u_1^*)],$$

which converges to 0 by the key renewal theorem. So there exists a $u_1' > 0$ such that when $u > u_1'$, the third term in (3.48) is bounded by $\frac{K\beta_e}{2\beta}\epsilon$. Let $l_1^* = \max(u_1', u_1^*)$. By (3.48) and summarizing the above, we obtain

$$\sup_{(\xi,\mu)\in\mathscr{I}_\epsilon^+} |x(u) - x(\infty)| < \frac{K\beta_e}{\beta}\epsilon \quad \text{for all } u > l_1^*.$$

Next, we consider the set $\mathscr{I}_\epsilon^-$. By doing the same algebra as in the proof of Lemma 3.8, we see that (3.49) holds for any $(\xi, \mu) \in \mathscr{I}_\epsilon^-$. By Lemma B.1, there exists a u_2^* such that

$$\sup_{(\xi,\mu)\in\mathscr{I}_\epsilon^-} |h_{\xi,\mu} * U_e(u) - \frac{\langle \chi, \xi+\mu\rangle}{\beta_e}| < \frac{K}{2}\epsilon.$$

for all $u > u_2^*$. So we have

$$h_{\xi,\mu} * U_e(u) \leq \frac{K\beta_e(1-\epsilon)}{\beta_e} + \frac{K}{2}\epsilon < K,$$

for all $(\xi,\mu) \in \mathscr{I}_\epsilon^-$ and $u > u_2^*$. Note that the last term in (3.49) is always non-positive. So when $u \geq u_2^*$ we have $x(u) < K$, (or equivalently $(x(u)-K)^+ = 0$) for all $(\xi,\mu) \in \mathscr{I}_\epsilon^-$. So by (3.44), the absolute value of the last term in (3.49) can be bounded by

$$\int_{u-u_2^*}^{u} \frac{w}{\beta} d[(1-F) * U_e(v)] = \frac{w}{\beta}[(1-F) * U_e(u) - (1-F) * U_e(u-u_2^*)],$$

which converges to 0 by the key renewal theorem. So there exists a $u_2' > 0$ such that when $u > u_2'$, the third term in (3.49) is bounded by $\frac{K}{2}\epsilon$. Let $l_2^* = \max(u_2', u_2^*)$. By (3.49) and summarizing the above,

$$\sup_{(\xi,\mu)\in\mathscr{I}_\epsilon^-} |x(u) - x(\infty)| < K\epsilon \quad \text{for all } u > l_2^*.$$

It only remains to deal with the set $\mathscr{I}_\epsilon^0$. For any $h_{\xi,\mu} \in \mathscr{I}_\epsilon^0$, consider the 2ϵ-perturbation $(\xi_{2\epsilon}, \mu_{2\epsilon}) \in \mathscr{I}_\epsilon^+$, and -2ϵ-perturbation $(\xi_{-2\epsilon}, \mu_{-2\epsilon}) \in \mathscr{I}_\epsilon^-$. Denote $x^+(\cdot)$ and $x^-(\cdot)$ the solutions to (3.23) corresponding to $(\xi_{2\epsilon}, \mu_{2\epsilon})$ and $(\xi_{-2\epsilon}, \mu_{-2\epsilon})$ respectively. By Lemma 3.7,

$$x^-(u) < x(u) < x^+(u) \quad \text{for all } u \geq 0.$$

According to the above two cases, when $l > l^* = \max(l_1^*, l_2^*)$,

$$x(u) \leq x^+(\infty) + \frac{K\beta_e}{\beta}\epsilon \leq x(\infty)(1+2\epsilon) + \frac{K\beta_e}{\beta}\epsilon,$$
$$x(u) \geq x^-(\infty) - K\epsilon \geq x(\infty)(1-2\epsilon) - K\epsilon,$$

for all $(\xi,\mu) \in \mathscr{I}_\epsilon^0$. This means that

$$\sup_{(\xi,\mu)\in\mathscr{I}_\epsilon^0} |x(u) - x(\infty)| < C\epsilon \quad \text{for all } u > l^*,$$

where $C = \max(2x(\infty) + \frac{K\beta_e}{\beta}, 2x(\infty) + K)$. □

Lemma 3.11. *Assume* (3.15)–(3.41) *and* (3.40). *For all* $\epsilon > 0$ *there exists an* $L^* > 0$ *such that when* $t \geq L^*$,

$$\sup_{(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\in\Xi(\mathscr{I}_M^p)} \mathbf{d}[\bar{\mathcal{Q}}(t), \frac{(w - K\beta_e)^+}{\beta}\nu] < \epsilon, \tag{3.55}$$

$$\sup_{(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\in\Xi(\mathscr{I}_M^p)} \sup_{y\in[0,\infty)} |\bar{\mathcal{Z}}(t)(A_y) - \frac{w \wedge K\beta_e}{\beta_e}\nu_e(A_y)| < \epsilon. \tag{3.56}$$

Proof. The proof of this corollary is almost the same as the proof of Lemma 3.9. Just note that by Lemma 3.10, the t_1 in (3.52) and the t_2 in (3.54) are good for all $(\xi, \mu) \in \mathscr{I}_M^p$. With $L^* = \max(t_1, t_2)$, the result of this lemma immediately follows. □

Proof of Theorem 3.3, part II. Now we use Lemma 3.11 to show the uniform convergence result. By Lemma C.1, (3.56) implies that for all $\epsilon > 0$ there exists an L_1^* such that when $t \geq L_1^*$

$$\sup_{(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\in\Xi(\mathscr{I}_M^p)} \mathbf{d}[\bar{\mathcal{Z}}(t), \frac{w \wedge K\beta_e}{\beta_e}\nu_e] < \epsilon.$$

The uniform convergence follows from this and (3.55). □

CHAPTER IV

FUNCTIONAL LAW OF LARGE NUMBER LIMITS

The main motivation to study the fluid model is that it serves as the weak law of large number limit of the stochastic process described in Chapter 2. Consider a sequence of limited processor sharing queues indexed by r, where r increases to ∞ through a sequence in $(0, \infty)$. Each model is defined in the same way as in Chapter 2. To distinguish models with different indices, quantities of the rth model are accompanied by superscript r. Each model may be defined on a different probability space $(\Omega^r, \mathcal{F}^r, \mathbb{P}^r)$. Our results concern the asymptotic behavior of the descriptor under the *fluid* scaling, which is defined by

$$\bar{\mathcal{Q}}^r(t) = \frac{1}{r}\mathcal{Q}^r(rt), \quad \bar{\mathcal{Z}}^r(t) = \frac{1}{r}\mathcal{Z}^r(rt), \tag{4.1}$$

for all $t \geq 0$. We are also interested in fluid scaled versions of other quantities like the workload and queue length processes. Note that $\bar{Q}^r(\cdot)$, $\bar{Z}^r(\cdot)$ and $\bar{W}^r(\cdot)$ are actually functions of $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$, so the scaling for these quantities are defined as the functions of the corresponding scaling for $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$, i.e.

$$\bar{Q}^r(t) = \langle 1, \bar{\mathcal{Q}}^r(t)\rangle = \frac{1}{r}Q^r(rt), \tag{4.2}$$

$$\bar{Z}^r(t) = \langle 1, \bar{\mathcal{Z}}^r(t)\rangle = \frac{1}{r}Z^r(rt), \tag{4.3}$$

$$\bar{W}^r(t) = \langle \chi, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t)\rangle = \frac{1}{r}W^r(rt), \tag{4.4}$$

for all $t \geq 0$. Similarly we define the fluid scaling for cumulative service amount $S^r(s,t)$ to be

$$\bar{S}^r(s,t) = \int_s^t \psi(\bar{Z}^r(\tau))d\tau, \tag{4.5}$$

for $0 \leq s \leq t$. The fluid scaling for the external arrival process is defined as

$$\bar{E}^r(t) = \frac{1}{r}E^r(rt). \tag{4.6}$$

It follows from (2.3) that the scaling for $\bar{B}^r(\cdot)$ and should be defined by

$$\bar{B}^r(t) = \frac{1}{r} B^r(rt), \tag{4.7}$$

for all $t \geq 0$.

To establish results on convergence of the above sequence of stochastic processes, we need the following conditions, which are quite general and standard. We assume that the arrival processes satisfy

$$\bar{E}^r(\cdot) \Rightarrow \lambda \cdot \quad \text{as } r \to \infty, \tag{4.8}$$

where λ is a positive constant. The job size measures ν^r satisfy that as $r \to \infty$

$$\mathbf{d}[\nu^r, \nu] \to 0, \tag{4.9}$$

$$\langle \chi^{1+p}, \nu^r \rangle \to \langle \chi^{1+p}, \nu \rangle < \infty \quad \text{for some } p > 0, \tag{4.10}$$

where ν satisfies

$$\nu \text{ has no atoms.} \tag{4.11}$$

The law of large number scaling speeds up the processes r times, so we need to scale the sharing limit accordingly:

$$\lim_{r \to \infty} K^r / r \to K > 0. \tag{4.12}$$

Also, the following initial condition will be assumed:

$$(\bar{\mathcal{Q}}^r(0), \bar{\mathcal{Z}}^r(0)) \Rightarrow (\xi^*, \mu^*), \tag{4.13}$$

$$\langle \chi^{1+p}, \bar{\mathcal{Q}}^r(0) + \bar{\mathcal{Z}}^r(0) \rangle \Rightarrow \langle \chi^{1+p}, \xi^* + \mu^* \rangle, \tag{4.14}$$

where p is the same as in (4.10) and (ξ^*, μ^*) is a deterministic element in $\mathscr{I}$ and

$$\mu^* \text{ has no atoms.} \tag{4.15}$$

The following proposition is a well known result for a single server queue operating under a non-idling service discipline. Readers are referred to Section 5 in [25] for a proof.

Proposition 4.1. *Assume the sequence of LPS queues satisfies* (4.8)–(4.14). *As* $r \to \infty$, *we have*

$$\bar{W}^r(\cdot) \Rightarrow \bar{W}(\cdot),$$

where $\bar{W}(t) = (\langle \chi, \xi^* + \mu^* \rangle + (1-\rho)t)^+$ *for all* $t \geq 0$.

Since the LPS is also a non-idling service discipline, the above limit of the workload process still holds for our model.

However, the limiting of the job size process and many other performance processes as introduced above is far from clear. Our main result establishes the fluid limit of the measure-valued processes (Theorem 4.1), from which the fluid limit of many interested performance processes follows directly (Corollary 4.1).

Theorem 4.1. *If the sequence of limited processor sharing queues satisfies* (4.8)–(4.15), *then*

$$(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot)) \Rightarrow (\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \quad \text{as } r \to \infty,$$

where $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *is the unique solution to the fluid model* (K, λ, ν) *with initial condition* (ξ^*, μ^*).

Since all performance measures can be recovered from the descriptor $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$, we have the following corollary.

Corollary 4.1. *Assume the sequence of limited processor queues satisfies* (4.8)–(4.15). *As* $r \to \infty$, *we have*

$$\left(\bar{Q}^r(\cdot), \bar{Z}^r(\cdot), \bar{B}^r(\cdot), \bar{D}^r(\cdot)\right) \Rightarrow \left(\bar{Q}(\cdot), \bar{Z}(\cdot), \bar{B}(\cdot), \bar{D}(\cdot)\right),$$

where $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$, $\bar{B}(\cdot)$, $\bar{D}(\cdot)$ *are as defined in* (3.1)–(3.5).

Corollary 4.1 follows immediately from Theorem 4.1. So we omit the proof for brevity. We will prove Theorem 4.1 at the end of this chapter.

4.1 Relative Compactness

The objective of this section is to show the precompactness property, Theorem 4.2 below, for the fluid scaled processes $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$ defined in Section 3.1.

Consider the rth system. A fluid scaled version of stochastic dynamic equations (2.5) and (2.6) can be written as

$$\bar{\mathcal{Q}}^r(t)(A') = \frac{1}{r} \sum_{i=r\bar{B}^r(t)+1}^{r\bar{E}^r(t)} \delta_{v_i^r}(A'),$$

$$\bar{\mathcal{Z}}^r(t)(A) = \frac{1}{r} \sum_{i=-r\bar{X}^r(0)+1}^{-r\bar{Q}^r(0)} \delta_{\tilde{v}_i^r}(A + \bar{S}^r(t)) + \frac{1}{r} \sum_{i=-r\bar{Q}^r(0)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}(A + \bar{S}^r(\tau_i, t)),$$

for $t \geq 0$ and any Borel sets $A' \subseteq [0, \infty)$ and $A \subseteq (0, \infty)$. Thus, by the above equations, we have for $0 \leq s \leq t$

$$\bar{\mathcal{Q}}^r(t)(A') = \bar{\mathcal{Q}}^r(s)(A') + \frac{1}{r} \sum_{i=r\bar{E}^r(s)+1}^{r\bar{E}^r(t)} \delta_{v_i^r}(A') - \frac{1}{r} \sum_{i=r\bar{B}^r(s)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}(A'), \tag{4.16}$$

$$\bar{\mathcal{Z}}^r(t)(A) = \bar{\mathcal{Z}}^r(s)(A + \bar{S}^r(s,t)) + \frac{1}{r} \sum_{i=r\bar{B}^r(s)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}(A + \bar{S}^r(\tau_i, t)). \tag{4.17}$$

The dynamics of the system is determined by the above equations. Equation (4.16) says that the status of the buffer at time t equals the status at time s plus what has arrived to the buffer and minus what has left from the buffer during time interval $(s, t]$. Those jobs who left buffer enter service, the service process has been taken care of by shifting the set A by the cumulative service amount $S^r(\tau_i, t)$ that the ith job receives. This corresponds to the second term on the right hand side of (4.17). This plus the status at time s shifted by accumulative service amount $S^r(s, t)$ is equal to the status of the server at time t, as indicated in (4.17). To simplify the notation in this section, for all $0 \leq s \leq t$, denote

$$\bar{E}^r(s,t) = \bar{E}^r(t) - \bar{E}^r(s), \quad \bar{B}^r(s,t) = \bar{B}^r(t) - \bar{B}^r(s).$$

Note that $\bar{\mathcal{Z}}^r(t) \in \mathbf{M}_2$ on each sample path for each $r > 0$ and $t > 0$. Due to the convention that $\mathbf{M}_2$ can be embedded in $\mathbf{M}_1$ (c.f. Section 1.4), we view $\bar{\mathcal{Z}}^r(t)$ as an

element in $\mathbf{M}_1$ when it is convenient. In particular, $\bar{\mathcal{Z}}^r(t)(A)$ is well defined for each Borel set $A \subset [0, \infty)$.

The compact containment property is derived in Section 4.1.1. Section 4.1.2 serves as a preparation for the oscillation bound. The oscillation bound is then proved in Section 4.1.3, followed by the precompactness result Theorem 4.2. The framework of the proofs is similar to that of [24, 26].

4.1.1 Compact Containment

The main objective of the section is to establish the compact containment property in Lemma 4.4, which is the first main step to prove precompactness. First, let us establish a bound for the arrival processes.

Fix $T > 0$. It follows immediately from condition (4.8) that for each $\epsilon, \epsilon' > 0$ there exists an r_0 such that when $r > r_0$,

$$\mathbb{P}\left(\sup_{0 \le s < t \le T} |\bar{E}^r(s,t) - \lambda(t-s)| < \epsilon'\right) \ge 1 - \epsilon. \tag{4.18}$$

To facilitate some arguments later on, we derive the following result from the above inequality.

Lemma 4.1. *Fix $T > 0$. There exists a function $\epsilon_E(\cdot)$, which vanishes at infinity such that*

$$\mathbb{P}\left(\sup_{0 \le s < t \le T} |\bar{E}^r(s,t) - \lambda(t-s)| < \epsilon_E(r)\right) \ge 1 - \epsilon_E(r),$$

for each $r \ge 0$.

Proof. For each index r let

$$H_r = \{\delta > 0 : (4.18) \text{ is true for } \epsilon' = \epsilon = \delta\}.$$

Clearly H_r is not empty since $1 \in H_r$. Let $\epsilon_E(r) = \inf H_r$ for each $r \ge 0$. Assume that $\epsilon_E(r)$ does not vanish at infinity. There exists a $\delta > 0$ and a sub-sequence $\{r_n\}_{n=1}^{\infty}$ which increases to infinity such that

$$\epsilon_E(r_n) \ge \delta \quad \text{for all } n \ge 0. \tag{4.19}$$

However, for $\epsilon' = \epsilon = \delta/2$ there exists an r_δ such that when $r_n \geq r_\delta$, (4.18) must hold. This contradicts (4.19). □

Denote

$$\Omega_E^r = \{ \sup_{t\in[0,T]} |\bar{E}^r(t) - \lambda t| < \epsilon_E(r)\}.$$

We have that

$$\lim_{r\to\infty} \mathbb{P}\left(\Omega_E^r\right) = 1. \tag{4.20}$$

It is clear from the policy constraint (2.8) that for all $t \geq 0$,

$$\bar{Z}^r(t) \leq K^r/r < K + 1, \tag{4.21}$$

where the last inequality holds for all large r since $K^r/r \to K$. The following lemma establishes a bound for the buffer size $\bar{Q}^r(\cdot)$.

Lemma 4.2. *Assume* (4.8) *and* (4.13). *Fix* $T > 0$. *For each* $\eta > 0$ *there exists a constant* $M_1 > 0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \bar{Q}^r(t) < M_1\right) \geq 1 - \eta.$$

Proof. Plugging $A = [0, \infty)$ in (4.16) and letting $s = 0$, we get

$$\bar{Q}^r(t) \leq \bar{Q}^r(0) + \bar{E}^r(t). \tag{4.22}$$

By condition (4.13), there exists a constant M' such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\bar{Q}^r(0) < M'\right) \geq 1 - \eta.$$

By (4.20) and (4.22), we have that

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \bar{Q}^r(t) < M' + \lambda T + 1\right) \geq 1 - \eta.$$

The lemma is proved by letting $M_1 = M' + \lambda T + 1$. □

Lemma 4.3. *Assume* (4.8)–(4.15). *Fix* $T > 0$. *For any* $\eta > 0$, *there exists a constant* $M_2 > 0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t) \rangle < M_2 \right) > 1 - \eta,$$

where the positive constant p *is the same as in conditions* (4.10) *and* (4.14).

Proof. By condition (4.14),

$$\liminf_{r\to\infty} \mathbb{P}\left(\langle \chi^{1+p}, \bar{\mathcal{Z}}^r(0) \rangle < \langle \chi^{1+p}, \xi^* + \mu^* \rangle + 1 \right) = 1.$$

Denote the event in the above by Ω_0^r. By Lemma 4.2, for any $\eta > 0$, there exists a constant $M_1 > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \bar{Q}^r(t) < M_1 \right) > 1 - \eta.$$

Denote the event in the above by $\Omega_1^r(M_1)$. Note that on the event $\Omega_1^r(M_1) \cap \Omega_E^r$,

$$\langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t) \rangle \le \langle \chi^{1+p}, \bar{\mathcal{Z}}^r(0) \rangle + \frac{1}{r} \sum_{i=-rM_1}^{\lfloor \lambda r T + r\epsilon_E(r) \rfloor} \langle \chi^{1+p}, \delta_{v_i^r} \rangle, \tag{4.23}$$

for any $t \in [0, T]$. By condition (4.10), $\langle \chi^{1+p}, \nu^r \rangle < \infty$ and $\langle \chi^{1+p}, \nu \rangle < \infty$. Since we only need to consider large enough r such that $\epsilon_E(r) < 1$, by Lemma A.2 in [25],

$$\liminf_{r\to\infty} \mathbb{P}\left(\frac{1}{r} \sum_{i=-rM_1}^{\lfloor \lambda r T + r\epsilon_E(r) \rfloor} \langle \chi^{1+p}, \delta_{v_i^r} \rangle < (\lambda T + M_1 + 1) \langle \chi^{1+p}, \nu \rangle + 1 \right) = 1.$$

Denote the above event by $\Omega_p^r(M_1)$, then by (4.20), we have

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_E^r \cap \Omega_0^r \cap \Omega_1^r(M_1) \cap \Omega_p^r(M_1) \right) > 1 - \eta. \tag{4.24}$$

The lemma is proved by letting $M_2 = \langle \chi^{1+p}, \xi^* + \mu^* \rangle + (\lambda T + M_1 + 1) \langle \chi^{1+p}, \nu \rangle + 2$. □

Denote

$$\Omega_B^r(M) = \left\{ \sup_{t\in[0,T]} \bar{Q}^r(t) < M \text{ and } \sup_{t\in[0,T]} \bar{Z}^r(t) < M \right\}$$
$$\cap \left\{ \sup_{t\in[0,T]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t) \rangle < M \right\}.$$

By (4.21), Lemmas 4.2 and 4.3, for any $\eta > 0$, there exists a constant $M > K + 1$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_B^r(M)\right) > 1 - \eta. \tag{4.25}$$

A set $\mathbf{K} \subset \mathbf{M}_1$ is relatively compact if $\sup_{\xi\in\mathbf{K}} \xi(\mathbb{R}_+) < \infty$, and if there exists a sequence of nested compact sets $J_n \subset \mathbb{R}_+$ such that $\bigcup J_n = \mathbb{R}_+$ and

$$\lim_{n\to\infty} \sup_{\xi\in\mathbf{K}} \xi(J_n^c) = 0,$$

where J_n^c denotes the complement of J_n; see [33], Theorem A7.5. Denote

$$\mathbf{K}(M) = \Big\{\xi \in \mathbf{M}_1 : \xi(\mathbb{R}_+) < M \text{ and } \xi((n,\infty)) \le M/n \text{ for all } n \in \mathbb{Z}_+\Big\}.$$

Clearly, $\mathbf{K}(M)$ is a relatively compact set for any constant $M > 0$.

Lemma 4.4. *On the event* $\Omega_B^r(M)$,

$$\bar{\mathcal{Q}}^r(t) \in \mathbf{K}(M) \text{ and } \bar{\mathcal{Z}}^r(t) \in \mathbf{K}(M) \text{ for all } t \in [0,T]$$

Proof. Note that both $\sup_{t\in[0,T]} \bar{\mathcal{Q}}^r(t)([0,\infty))$ and $\sup_{t\in[0,T]} \bar{\mathcal{Z}}^r(t)((0,\infty))$ are bounded by M, according to the definition of $\Omega_B^r(M)$. By Markov's inequality, for any $t \ge 0$,

$$\bar{\mathcal{Q}}^r(t)((n,\infty)) \le \frac{\langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t)\rangle}{n^{1+p}},$$

which is bounded by $\frac{M}{n^{1+p}}$ by the definition of $\Omega_B^r(M)$. The same argument applies for $\bar{\mathcal{Z}}^r(t)$. □

4.1.2 Asymptotic Regularity

The second major step to prove precompactness is to obtain the oscillation bound in Section 4.1.3. Oscillations mainly result from sudden departures of a large number of jobs. To control the departure process, we show that $\bar{\mathcal{Z}}^r(\cdot)$ assigns arbitrarily small mass to small intervals. Similar results have been proved for PS queues and related models, see [24, 26]. In our model, the process of jobs entering the service is $\bar{B}^r(t) = \bar{E}^r(t) - \bar{Q}^r(t)$ instead of $\bar{E}^r(t)$, which creates additional difficulties.

Recall the Glivenko-Cantelli estimate in Lemma D.2. By the same argument as in Lemma 4.1, for fixed $M, T > 0$, there exists a function $\epsilon_{\mathrm{GC}}(\cdot)$, which vanishes at infinity, such that the probability inequality in Lemma D.2 holds with ϵ and ϵ' replaced by this function. In other words, denote

$$\Omega^r_{\mathrm{GC}}(M) = \Big\{ \max_{-rM<n<r(M+2\lambda T)} \sup_{l\in[0,2M+2\lambda T]} \sup_{f\in\overline{\mathscr{V}}} |\langle f, \bar{\eta}^r(n,l)\rangle - l\langle f, \nu^r\rangle| < \epsilon_{\mathrm{GC}}(r) \Big\},$$

where

$$\bar{\eta}^r(n,l) = \frac{1}{r}\sum_{i=n+1}^{n+\lfloor rl\rfloor} \delta_{v_i^r},$$

and $\overline{\mathscr{V}}$ is a set of functions of the form $1_{(x,\infty)}$ and $1_{[x,\infty)}$ for all $x \in \mathbb{R}_+$ with an envelope function $\bar{f}$ (see Section D). We have

$$\lim_{r\to\infty} \mathbb{P}\left(\Omega^r_{\mathrm{GC}}(M)\right) = 1. \tag{4.26}$$

The Glivenko-Cantelli estimate helps prove the following result.

Lemma 4.5. *Assume* (4.8)–(4.15). *Fix* $T > 0$. *For each* $\epsilon, \eta > 0$ *there exists a* $\kappa > 0$ *(depending on* ϵ *and* η*) such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathscr{Z}}^r(t)([x, x+\kappa]) \le \epsilon \right) \ge 1-\eta. \tag{4.27}$$

Proof. We first show that for any $\epsilon, \eta > 0$, there exists a κ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}_+} \bar{\mathscr{Z}}^r(0)([x, x+\kappa]) \le \epsilon/2 \right) \ge 1-\eta/2. \tag{4.28}$$

It follows from the initial condition (4.13) that $\bar{\mathscr{Z}}^r(0) \Rightarrow \mu^*$ as $r \to \infty$. Since μ^* is a finite Borel measure, there exists an $M > 0$ such that

$$\mu^*([M,\infty)) < \epsilon/4.$$

By (4.15), the distribution function associated with the measure μ^* is continuous, thus is uniformly continuous on the finite interval $[0, 2M]$. So there exists a $\kappa \in (0, M]$ such that

$$\sup_{x\in[0,M]} \mu^*([x, x+\kappa]) < \epsilon/4.$$

The above two inequalities imply

$$\sup_{x\in\mathbb{R}_+} \mu^*([x, x+\kappa]) < \epsilon/4.$$

Let $N = \lceil M/\kappa \rceil$. Denote $I_n = [n\kappa, (n+2)\kappa]$ for $n = 0, 1, \ldots, N-1$, and $I_N = [M, \infty)$. Note that, for every $x \in [0, \infty)$ there exists an $n \leq N$ such that $[x, x+\kappa] \subset I_n$. To prove (4.28), it suffices to show

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{n\leq N} \bar{\mathcal{Z}}^r(0)(I_n) \leq \epsilon/2\right) \geq 1 - \eta/2. \tag{4.29}$$

Denote $\mathbf{A} = \{\mu \in \mathbf{M}_2 : \max_{n\leq N} \mu(I_n) < \epsilon/2\}$. It is clear that $\mu^* \in \mathbf{A}$. Now, let us prove that the set $\mathbf{A}$ is open in the space $\mathbf{M}_2$ equipped with the Prohorov metric. Let $\{\mu_k\} \subset \mathbf{M}_2$ be a sequence in the Polish space $\mathbf{M}_2$ satisfying $\mu_k \to \mu$ for some $\mu \in \mathbf{A}$. Since each I_n is closed, by the Portmanteau theorem, Theorem 2.1 in [5] (adapted to finite measures, see also [26]),

$$\limsup_{k\to\infty} \mu_k(I_n) \leq \mu(I_n) < \epsilon/2 \quad \text{for all } n \leq N.$$

Hence, $\mu_k \in \mathbf{A}$ for all sufficiently large k, which implies that $\mathbf{A}$ is open in $\mathbf{M}$. Thus, a second application of the Portmanteau theorem yields

$$\liminf_{r\to\infty} \mathbb{P}\left(\bar{Z}^r(0) \in \mathbf{A}\right) \geq \mathbb{P}\Big(\mu^* \in \mathbf{A}\Big) = 1,$$

which implies (4.29).

Now we need to extend this result to the interval $[0, T]$. Denote the event in (4.28) by Ω_1^r. Let

$$\Omega_2^r(M) = \Omega_1^r \cap \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M).$$

By (4.20), (4.25) and (4.26), there exists an $M > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_2^r(M)\right) \geq 1 - \eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega_2^r(M)$.

For any $r > 0$, $t \in [0, T]$ we define the random time

$$t_0 = \sup\Big\{\{s \le t : \langle 1, \bar{\mathcal{Z}}^r(s)\rangle < \epsilon/4\} \cup \{0\}\Big\}.$$

If $t_0 = 0$, then by (4.28) for each $x \in \mathbb{R}_+$

$$\bar{\mathcal{Z}}^r(0)([x, x+\kappa] + \bar{S}^r(t)) \le \epsilon/2.$$

If $t_0 \in (0, t]$, then for each $\delta > 0$ there exists an s such that $t_0 - \delta < s < t_0$ and $\bar{Z}^r(s)(\mathbb{R}_+) < \epsilon/4$. Since we are only concerned with small ϵ (which should be small enough such that $\bar{Z}^r(s) < \epsilon/4 < K^r/r$), $\bar{Q}^r(s) = 0$ by the policy constraint (2.8). Note that (2.3) implies

$$\bar{B}^r(s', t) \le \bar{E}^r(s', t) + \bar{Q}^r(s') \text{ for all } s' \le t. \tag{4.30}$$

Since we are on the event Ω_E^r, for any $\epsilon_1 > 0$, we have $\bar{B}^r(s, t_0) \le \lambda\delta + \epsilon_1$ for all large enough r. For any Borel set A, by the fluid scaled system dynamic equation (4.17),

$$\bar{\mathcal{Z}}^r(t_0)(A) \le \bar{\mathcal{Z}}^r(s)(\mathbb{R}_+) + \bar{B}^r(s, t_0) \le \epsilon/4 + \lambda\delta + \epsilon_1,$$

which can be made smaller than $\epsilon/2$ by choosing ϵ_1, δ suitably small.

The fluid scaled stochastic dynamic equation over the interval $[t_0, t]$ can be written as

$$\begin{aligned}\bar{\mathcal{Z}}^r(t)([x, x+\kappa]) = \bar{\mathcal{Z}}^r(t_0)([x, x+\kappa] + \bar{S}^r(t_0, t)) \\ + \frac{1}{r}\sum_{i=r\bar{B}^r(t_0)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}([x, x+\kappa] + \bar{S}^r(\tau_i, t)),\end{aligned}$$

for each $x \in \mathbb{R}_+$. By the choice of t_0, the first term on the right hand side of the above equation is always upper bounded by $\epsilon/2$. Let I denote the second term on the right hand side of the proceeding equation. Now it only remains to show that $I < \epsilon/2$.

Let $t_0, t_1, \cdots, t_N = t$ be a partition of the interval $[t_0, t]$ such that $|t_{j+1} - t_j| < \delta$ for all $j = 0, \cdots, N-1$, where δ and N are to be chosen below. Write I as the

summation

$$I = \sum_{j=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{B}^r(t_j)+1}^{r\bar{B}^r(t_{j+1})} \delta_{v_i^r}([x, x+\kappa] + \bar{S}^r(\tau_i, t)).$$

Recall that τ_i^r is the time that the ith job starts service, so on each sub-interval $[t_j, t_{j+1}]$ those i's to be summed must satisfy $t_j \le \tau_i^r \le t_{j+1}$. This implies that

$$\bar{S}^r(t_{j+1}, t) \le \bar{S}^r(\tau_i, t) \le \bar{S}^r(t_j, t).$$

By the definition of t_0, we have $\bar{Z}^r(s) \ge \epsilon/4$ for all $s \in [t_0, t]$. So

$$\bar{S}^r(t_j, t_{j+1}) \le \frac{4\delta}{\epsilon}.$$

Let

$$C_j = [x + \bar{S}^r(t_{j+1}, t), \quad x + \bar{S}^r(t_{j+1}, t) + \kappa + \frac{4\delta}{\epsilon}].$$

Then

$$I \le \sum_{j=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{B}^r(t_j)+1}^{r\bar{B}^r(t_{j+1})} \delta_{v_i^r}(C_j).$$

Since we are on the event $\Omega_E^r \cap \Omega_B^r(M)$, by (4.30), we have for all $j = 0, \cdots, N-1$

$$-rM \le r\bar{B}^r(t_j) \le r(\lambda T + \epsilon_1 + M) \le 2\lambda r T + rM,$$

$$\bar{B}^r(t_j, t_{j+1}) \le \lambda T + \epsilon_1 + M \le 2\lambda T + M.$$

Since we are on the event $\Omega_{\mathrm{GC}}^r(M)$,

$$\Big|\frac{1}{r} \sum_{i=r\bar{B}^r(t_j)+1}^{r\bar{B}^r(t_{j+1})} \delta_{v_i^r}(C_j) - \Big(\bar{B}^r(t_{j+1}) - \bar{B}^r(t_j)\Big)\nu^r(C_j)\Big| < \epsilon_1.$$

So

$$I \le \sum_{j=0}^{N-1} [\bar{B}^r(t_{j+1}) - \bar{B}^r(t_j)]\nu^r(C_j) + N\epsilon_1.$$

By (4.9), for all $\epsilon_2 > 0$

$$d[\nu^r, \nu] \le \epsilon_2,$$

for all large enough r. Note that C_j is a closed Borel set, by the definition of Prohorov metric, we have

$$\nu^r(C_j) \le \nu(C_j^{\epsilon_2}) + \epsilon_2$$

for all large enough r. Since $C_j^{\epsilon_2}$ is a closed interval with length $\kappa + \frac{4\delta}{\epsilon} + 2\epsilon_2$, by (4.11) we can choose $\kappa, \delta, \epsilon_2$ small enough such that

$$\nu(C_j^{\epsilon_2}) < \frac{\epsilon}{4(2\lambda T + M)}.$$

Thus, we conclude that

$$\begin{aligned} I &\leq (\epsilon_2 + \frac{\epsilon}{4(2\lambda T + M)}) \sum_{j=0}^{N-1} [\bar{B}^r(t_{j+1}) - \bar{B}^r(t_j)] + N\epsilon_1 \\ &\leq (\epsilon_2 + \frac{\epsilon}{4(2\lambda T + M)})[\bar{B}^r(t) - \bar{B}^r(t_0)] + N\epsilon_1 \\ &\leq \epsilon_2(2\lambda T + M) + \epsilon/4 + N\epsilon_1, \end{aligned}$$

where the last inequality is because we are on the event $\Omega_E^r \cap \Omega_B^r(M)$. Finally, by choosing ϵ_1, ϵ_2 small enough, we obtain that $I < \epsilon/2$. □

In addition to the asymptotic regularity for the server $\bar{\mathcal{Z}}^r(\cdot)$, we also have the same property for the buffer $\bar{\mathcal{Q}}^r(\cdot)$. The proof is much easier.

Lemma 4.6. *Assume* (4.8)–(4.15). *Fix* $T > 0$. *For each* $\epsilon, \eta > 0$ *there exists a* $\kappa > 0$ *(depending on* ϵ *and* η*) such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Q}}^r(t)([x, x+\kappa]) \leq \epsilon \right) \geq 1 - \eta. \tag{4.31}$$

Proof. Let

$$\Omega_3^r(M) = \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M).$$

By (4.20), (4.25) and (4.26), there exists an $M > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_3^r(M)\right) \geq 1 - \eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega_3^r(M)$.

Since we are on the event $\Omega_E^r \cap \Omega_B^r(M)$, $|\bar{B}^r(\cdot)|$ and $\bar{E}^r(\cdot)$ are bounded above by $M + 2\lambda T$. Since we are on the event $\Omega_{\mathrm{GC}}^r(M)$, for any $t \in [0,T]$ and $\epsilon_1 > 0$,

$$\begin{aligned}
&\left|\bar{\mathcal{Q}}^r(t)([x,x+\kappa]) - \big(\bar{E}^r(t) - \bar{B}^r(t)\big)\nu^r([x,x+\kappa])\right| \\
&= \left|\frac{1}{r}\sum_{i=r\bar{B}^r(t)+1}^{r\bar{E}^r(t)} \delta_{v_i^r}([x,x+\kappa]) - \big(\bar{E}^r(t) - \bar{B}^r(t)\big)\nu^r([x,x+\kappa])\right| \\
&\leq \epsilon_1,
\end{aligned}$$

for all large r. Thus

$$\begin{aligned}
\bar{\mathcal{Q}}^r(t)([x,x+\kappa]) &\leq \big(\bar{E}^r(t) - \bar{B}^r(t)\big)\nu^r([x,x+\kappa]) + \epsilon_1 \\
&\leq 2M\nu^r([x,x+\kappa]) + \epsilon_1,
\end{aligned}$$

for all large r. By (4.9), for any $\epsilon_2 > 0$,

$$d[\nu^r, \nu] \leq \epsilon_2,$$

for all large enough r. By the definition of Prohorov metric, we have

$$\nu^r([x,x+\kappa]) \leq \nu([x-\epsilon_2, x+\kappa+\epsilon_2]) + \epsilon_2$$

for all large enough r. By (4.11), we can choose κ, ϵ_2 small enough such that

$$\nu([x-\epsilon_2, x+\kappa+\epsilon_2]) < \epsilon_1.$$

Thus, we conclude that for any $t \in [0,T]$,

$$\bar{\mathcal{Q}}^r(t)[x,x+\kappa] \leq 2M(\epsilon_1+\epsilon_2) + \epsilon_1.$$

The proof is completed by choosing ϵ_1 and ϵ_2 to be less than $\epsilon/8M$. □

4.1.3 Oscillation Bound

In this section, we use the regularity result in Lemma 4.5 to obtain the oscillation bound in Lemma 4.7. The proof technique of this lemma is a simplification of that for

Lemma 4.14 in [24]. Consider a *càdlàg* function $\zeta(\cdot)$ on a fixed interval $[0,T]$ taking values in a metric space $(\mathbf{E},\pi)$. For $T \geq 0$ and $\delta > 0$, define the *modulus of continuity* to be

$$\mathbf{w}_L(\zeta(\cdot),\delta) = \sup_{s,t\in[0,T],|s-t|<\delta} \pi[\zeta(s),\zeta(t)].$$

If the metric space is $\mathbb{R}$, we just use the Euclidean metric; if the space is $\mathbf{M}_1 \times \mathbf{M}_2$, we use the Prohorov metric $\mathbf{d}$ defined in Section 1.

Lemma 4.7. *Assume* (4.8)–(4.15). *Fix* $T > 0$. *For each* $\epsilon, \eta > 0$ *there exists a* $\delta > 0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\max\left(\mathbf{w}_L(\bar{\mathcal{Q}}^r(\cdot),\delta), \mathbf{w}_L(\bar{\mathcal{Z}}^r(\cdot),\delta)\right) \leq \epsilon\right) \geq 1-\eta. \tag{4.32}$$

Proof. For any $\kappa, \epsilon > 0$, define

$$\Omega^r_{\mathrm{Reg}}(\kappa,\epsilon) = \Big\{ \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^r(t)([x,x+\kappa]) \leq \epsilon/5 \Big\}.$$

By (4.20) and Lemma 4.5, for each $\epsilon > 0$ and $\eta > 0$ there exists a $\kappa > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega^r_E \cap \Omega^r_{\mathrm{Reg}}(\kappa,\epsilon)\right) > 1-\eta.$$

In the remainder of the proof, we set

$$\delta = \min\left(\epsilon/5\lambda, \kappa\epsilon/5, \epsilon^2/25\right)$$

and all random quantities with index r are evaluated at a fixed sample path $\omega \in \Omega^r_E \cap \Omega^r_{\mathrm{Reg}}(\kappa,\epsilon)$. For $0 \leq s \leq t \leq T$ with $t - s < \delta$, consider the following two cases.

Case 1 If $\inf_{\tau\in[s,t]} \bar{X}^r(\tau) < \epsilon/5$, let

$$t_0 = \inf\{\tau \in [s,t] : \bar{X}^r(\tau) \leq \epsilon/5\}.$$

By right continuity, $\bar{X}^r(t_0) \leq \epsilon/5$. We only need consider large enough r's such that $\epsilon/5$ is smaller than K^r/r (which converges to $K > 0$ as $r \to \infty$ by condition (4.12)).

On the interval $[s, t_0)$, $\bar{Z}^r(\cdot)$ is larger than $\epsilon/5$, implying $\bar{S}^r(s, t_0) \leq \frac{|t_0 - s|}{\epsilon/5} \leq \frac{|t-s|}{\epsilon/5}$. So we have

$$\bar{\mathcal{Z}}^r(s)(A_0 + \frac{|t-s|}{\epsilon/5}) \leq \bar{\mathcal{Z}}^r(s)(A_0 + \bar{S}^r(s, t_0)) \leq \bar{\mathcal{Z}}^r(t_0)(A_0) \leq \bar{X}^r(t_0) \leq \epsilon/5, \quad (4.33)$$

where $A_0 = (0, \infty)$ and the second inequality is due to (4.17). Note that $\delta \leq \kappa\epsilon/5$ implies that $\frac{|t-s|}{\epsilon/5} < \kappa$. Thus

$$\bar{Z}^r(s) = \bar{\mathcal{Z}}^r(s)(A_0) \leq \bar{\mathcal{Z}}^r(s)\Big(A_0 + \frac{|t-s|}{\epsilon/5}\Big) + \bar{\mathcal{Z}}^r(s)((0, \kappa]) \leq 2\epsilon/5,$$

where the last inequality follows from (4.33) and the definition of $\Omega^r_{\text{Reg}}(\kappa, \epsilon)$. So we have

$$\bar{\mathcal{Q}}^r(s) = \mathbf{0}, \quad \mathbf{d}[\bar{\mathcal{Z}}^r(s), \mathbf{0}] \leq 2\epsilon/5.$$

On the other hand, we have

$$\bar{X}^r(t) \leq \bar{X}^r(s) + \bar{E}^r(s, t) \quad \text{for all } s \leq t.$$

Since we are on the event Ω^r_E and we can choose r large enough such that $\epsilon_E(r) < \epsilon/5$, we have

$$\bar{E}^r(s, t) \leq \lambda\delta + \epsilon/5 \leq 2\epsilon/5, \quad (4.34)$$

where the last inequality is due to the choice of δ. So $\bar{X}^r(t) \leq \bar{X}^r(t_0) + 2\epsilon/5 = 3\epsilon/5$. Again, we only need consider large enough r's such that $\epsilon/5 + 2\epsilon/5 < K^r/r$. So we have

$$\bar{\mathcal{Q}}^r(t) = \mathbf{0}, \quad \mathbf{d}[\bar{\mathcal{Z}}^r(t), \mathbf{0}] \leq 3\epsilon/5.$$

In summary, we have that when $|t - s| \leq \delta$,

$$\mathbf{d}[\bar{\mathcal{Q}}^r(s), \bar{\mathcal{Q}}^r(t)] = 0, \quad \mathbf{d}[\bar{\mathcal{Z}}^r(s), \bar{\mathcal{Z}}^r(t)] \leq 3\epsilon/5 + 2\epsilon/5 = \epsilon.$$

Case 2 If $\inf_{\tau \in [s,t]} \bar{X}^r(\tau) \geq \epsilon/5$, then $\inf_{\tau \in [s,t]} \bar{Z}^r(\tau) \geq \epsilon/5$. Therefore,

$$\bar{S}^r(s, t) \leq \frac{t-s}{\epsilon/5} \leq \frac{\delta}{\epsilon/5} \leq \min(\kappa, \epsilon/5), \quad (4.35)$$

by the choice of δ. The number of jobs that enter the service during time interval $(s,t]$ is

$$\bar{B}^r(s,t) \le \bar{E}^r(s,t) + \bar{\mathcal{Z}}^r(s)([0,\bar{S}^r(s,t)]) \le 3\epsilon/5, \tag{4.36}$$

by (4.34), the choice of δ and the definition of Ω^r_{Reg}. By the dynamic equation (4.16), we have

$$|\bar{\mathcal{Q}}^r(s)(A) - \bar{\mathcal{Q}}^r(t)(A)| \le \max\left(\bar{E}^r(s,t), \bar{B}^r(s,t)\right) \le 3\epsilon/5$$

for any Borel set A. Thus

$$\mathbf{d}[\bar{\mathcal{Q}}^r(s), \bar{\mathcal{Q}}^r(t)] \le 3\epsilon/5.$$

By the dynamic equation (4.17),

$$\bar{\mathcal{Z}}^r(t)(A) \le \bar{\mathcal{Z}}^r(s)(A + \bar{S}^r(s,t)) + \bar{B}^r(s,t).$$

By (4.35), $A + \bar{S}^r(s,t) \subset A^{3\epsilon/5}$, where A^a is the a-enlargement of the set A as defined in Section 1.4. So by (4.36)

$$\bar{\mathcal{Z}}^r(t)(A) \le \bar{\mathcal{Z}}^r(s)(A^{3\epsilon/5}) + 3\epsilon/5 \quad \text{for any Borel set } A.$$

By Property (ii) on page 72 in [5], we have $\mathbf{d}[\bar{\mathcal{Z}}^r(s), \bar{\mathcal{Z}}^r(t)] \le 3\epsilon/5$. □

For any sequences $\{\kappa_i\}$ and $\{\delta_i\}$ of positive numbers, consider the following set

$$\begin{aligned}
&\Big\{ \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Q}}^r(t)([x,x+\kappa_j]) \le \frac{1}{j} \Big\} \\
\cap &\Big\{ \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^r(t)([x,x+\kappa_j]) \le \frac{1}{j} \Big\} \\
\cap &\{\max\left(\mathbf{w}_L(\bar{\mathcal{Q}}^r(\cdot),\delta_j), \mathbf{w}_L(\bar{\mathcal{Z}}^r(\cdot),\delta_j)\right) \le \frac{1}{j}\Big\}.
\end{aligned}$$

Denote the two sequences $\{\kappa_j\}$ and $\{\delta_j\}$ by $\mathcal{S}$. To emphasize the dependency on $\mathcal{S}$ and j, denote the above event by $\Omega^r_R(\mathcal{S},j)$. By Lemmas 4.5, 4.6 and 4.7, for any $\eta > 0$, there exists an $\mathcal{S}$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega^r_R(\mathcal{S},j)\right) \ge 1 - \frac{\eta/2}{2^j} \quad \text{for each } j \in \mathbb{N}.$$

For any finite number $n \in \mathbb{N}$, by the above inequality, we have

$$\liminf_{r\to\infty} \mathbb{P}\left(\bigcap_{j=1}^{n} \Omega_R^r(\mathcal{S}, j)\right) \geq 1 - \eta/2.$$

Let $r(n)$ denote the smallest number such that

$$\mathbb{P}\left(\bigcap_{j=1}^{n} \Omega_R^r(\mathcal{S}, j)\right) \geq 1 - \eta, \quad \text{for all } r \geq r(n). \tag{4.37}$$

It is clear that $r(\cdot)$ is a function defined on $\mathbb{Z}_+$ and it is non-decreasing (since $\bigcap_{j=1}^{n} \Omega_R^r(\mathcal{S}, j) \subset \bigcap_{j=1}^{n'} \Omega_R^r(\mathcal{S}, j)$ for any $n < n'$). Let

$$n(r) = \sup\Big\{\{n \in \mathbb{Z}_+ : r(n) \leq r\} \cup \{0\}\Big\}.$$

(From the definition, we see that $n(r)$ is allowed to be infinite. For example, when the function $r(\cdot)$ has an upper bound.) In fact, $n(\cdot)$ can be viewed as the "inverse" of $r(\cdot)$. It is clear that $n(\cdot)$ is an non-decreasing. We claim that $\lim_{r\to\infty} n(r) = \infty$. The reason is as follows: for any $n_0 > 0$ there exists $r_0 = r(n_0)$ such that $n(r) \geq n_0$ for all $r \geq r_0$. Now define

$$\Omega_R^r(\mathcal{S}) = \bigcap_{j=1}^{n(r)} \Omega_R^r(\mathcal{S}, j).$$

Note that $\Omega_R^r(\mathcal{S})$ is not empty for all large enough r (since $n(r) > 1$ for all large enough r), and in this case,

$$\mathbb{P}\left(\Omega_R^r(\mathcal{S})\right) \geq 1 - \eta.$$

So we conclude that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_R^r(\mathcal{S})\right) \geq 1 - \eta. \tag{4.38}$$

Now, denote

$$\Omega^r(M, \mathcal{S}) = \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M) \cap \Omega_R^r(\mathcal{S}).$$

For any r, the rth system is defined on the probability space $(\Omega^r, \mathbb{P}^r, \mathcal{F}^r)$. The stochastic processes $\mathcal{Q}^r(\cdot)$ and $\mathcal{Z}^r(\cdot)$ are actually measurable functions on Ω^r. From now on, in some statements we will explicitly write them down in the form of $\mathcal{Q}^r(\omega, \cdot)$

and $\mathcal{Z}^r(\omega,\cdot)$ to indicate that they are evaluated at the sample path $\omega \in \Omega^r$. We are now ready to present the precompactness result.

Theorem 4.2. *Assume* (4.8)–(4.15). *Fix* $T > 0$. *For all* $\eta > 0$, *the exists a constant* $M > 0$ *and an* $\mathcal{S}$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega^r(M,\mathcal{S})\right) \geq 1-\eta. \tag{4.39}$$

Any sequence of functions $\{(\bar{\mathcal{Q}}^{r_n}(\omega^{r_n},\cdot),\bar{\mathcal{Z}}^{r_n}(\omega^{r_n},\cdot))\}_{n\in\mathbb{N}}$ *with* $\omega^{r_n} \in \Omega^{r_n}(M,\mathcal{S})$ *for each* $n \in \mathbb{N}$ *and* $\{r_n\}_{n\in\mathbb{N}}$ *increasing to infinity has a subsequence* $\{(\bar{\mathcal{Q}}^{r_{n_i}}(\omega^{r_{n_i}},\cdot),\bar{\mathcal{Z}}^{r_{n_i}}(\omega^{r_{n_i}},\cdot))\}_{i\in\mathbb{N}}$ *such that*

$$\sup_{t\in[0,T]} \mathbf{d}[(\bar{\mathcal{Q}}^{r_{n_i}}(\omega^{r_{n_i}},t),\bar{\mathcal{Z}}^{r_{n_i}}(\omega^{r_{n_i}},t)),(\tilde{\mathcal{Q}}(t),\tilde{\mathcal{Z}}(t))] \to 0 \quad \text{as } i\to\infty,$$

for some process $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))$ *which is continuous.*

Proof. The probability inequality follows immediately from (4.20), (4.25), (4.26) and (4.38).

The space $\mathbf{M}_1\times\mathbf{M}_2$ endowed with the metric $\mathbf{d}$ (defined in Section 1.4) is complete. Lemma 4.4 verifies condition (a) in Theorem 3.6.3 of [17]. For any $\epsilon > 0$ there exists a j_0 such that $1/j < \epsilon$ for all $j \geq j_0$. Since we are on the event $\Omega^r_R(\mathcal{S})$, we have that when $\delta \leq \delta_{j_0}$ and r is large enough such that $n(r) > j_0$,

$$\max\left(\mathbf{w}_T(\mathcal{Q}^r(\omega^r,\cdot),\delta),\mathbf{w}_T(\mathcal{Z}^r(\omega^r,\cdot),\delta)\right) < \epsilon, \tag{4.40}$$

for any $\omega^r \in \Omega^r(M,\mathcal{S})$. This verifies condition (b) in Theorem 3.6.3 of [17]. So the sequence $\{(\bar{\mathcal{Q}}^{r_n}(\omega^{r_n},\cdot),\bar{\mathcal{Z}}^{r_n}(\omega^{r_n},\cdot))\}_{n\in\mathbb{N}}$ is precompact in the space $\mathbf{D}([0,T],\mathbf{M}_1\times\mathbf{M}_2)$ endowed with the Skorohod J_1 topology. In other words, there is a convergent subsequence. The limit of this subsequence is continuous by the oscillation bound (4.40). So convergence in the Skorohod J_1 topology is the same as convergence in the uniform metric defined in Section 1.4. □

4.2 Characterization of Limits

Let $\mathscr{D}_T(M,\mathcal{S})$ denote the set of limits of all convergent subsequences of the sequences in Theorem 4.2. It is clear that $\mathscr{D}_T(M,\mathcal{S})$ is a non-empty subset of elements in the space $\mathbf{D}([0,T],\mathbf{M}_1\times\mathbf{M}_2)$. We have the following result (Theorem 4.3) about the set $\mathscr{D}_T(M,\mathcal{S})$. The proof of Theorem 4.1, which builds on this result, will be provided at the end of the section.

Theorem 4.3. *$\mathscr{D}_T(M,\mathcal{S})$ contains only one element, which is the unique fluid model solution $(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))$ with initial condition (ξ^*,μ^*) restricted on the interval $[0,T]$.*

To better structure the proof, we first present three auxiliary lemmas (Lemmas 4.8, 4.9 and 4.10), which characterize any fixed element $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))$ in the set $\mathscr{D}_T(M,\mathcal{S})$. By the definition of $\mathscr{D}_T(M,\mathcal{S})$, for any $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))\in\mathscr{D}_T(M,\mathcal{S})$, there exists a sequence $\{r_n\}$ which goes to ∞ and $\omega^{r_n}\in\Omega^{r_n}(M,\mathcal{S})$ for each r_n such that

$$\sup_{t\in[0,T]}\mathbf{d}[(\bar{\mathcal{Q}}^{r_n}(\omega^{r_n},t),\bar{\mathcal{Z}}^{r_n}(\omega^{r_n},t)),(\tilde{\mathcal{Q}}(t),\tilde{\mathcal{Z}}(t))]\to 0\quad\text{as } n\to\infty.$$

With a slight abuse of notation, we drop the parameter ω^{r_n} for simplicity in the proofs of all the following three lemmas. We then have

$$\lim_{n\to\infty}\sup_{t\in[0,T]}\mathbf{d}[\bar{\mathcal{Q}}^{r_n}(t),\tilde{\mathcal{Q}}(t)]=0, \tag{4.41}$$

$$\lim_{n\to\infty}\sup_{t\in[0,T]}\mathbf{d}[\bar{\mathcal{Z}}^{r_n}(t),\tilde{\mathcal{Z}}(t)]=0. \tag{4.42}$$

Lemma 4.8. *Assume (4.8)–(4.15). For any point $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))\in\mathscr{D}_T(M,\mathcal{S})$, both $\tilde{\mathcal{Q}}(t)$ and $\tilde{\mathcal{Z}}(t)$ are atom free for all $t\in[0,T]$.*

Proof. For any $y\geq 0$ and $\kappa_1>0$, since $[y-\kappa_1,y+2\kappa_1]$ is the κ_1-enlargement (c.f. Section 1.4) of the set $[y,y+\kappa_1]$, by (4.41) and the definition of Prohorov metric, we have

$$\tilde{\mathcal{Q}}(t)([y,y+\kappa_1])\leq\bar{\mathcal{Q}}^{r_n}(t)([y-\kappa_1,y+2\kappa_1])+\kappa_1,$$

for all large n. Since we are on the event $\Omega^{r_n}(M,\mathcal{S})$, in particular $\Omega_R^{r_n}(\mathcal{S})$, for any $\epsilon > 0$, we can choose κ_1 small enough such that

$$\bar{\mathcal{Q}}^{r_n}(t)([y-\kappa_1, y+2\kappa_1]) < \epsilon/2,$$

for all large n. When making κ_1 small, we can also choose $\kappa_1 < \epsilon/2$. This gives that

$$\tilde{\mathcal{Q}}(t)([y, y+\kappa_1]) < \epsilon.$$

This proves that $\tilde{\mathcal{Q}}(t)$ is atom free for any $t \in [0,T]$. The proof for $\tilde{\mathcal{Z}}(t)$ follows in exactly the same way. □

Lemma 4.9. *Assume* (4.8)–(4.15). *Fix any point* $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_T(M,\mathcal{S})$ *and constants* $a, b \in [0,T]$ *with* $a < b$. *If*

$$\inf_{t\in[a,b]} \tilde{Z}(t) > 0, \tag{4.43}$$

then $(\tilde{\mathcal{Q}}(a), \tilde{\mathcal{Z}}(a)) \in \mathscr{I}$ *and* $(\tilde{\mathcal{Q}}(a+\cdot), \tilde{\mathcal{Z}}(a+\cdot))$ *is the solution to the fluid model* (K,λ,ν) *with initial condition* $(\tilde{\mathcal{Q}}(a), \tilde{\mathcal{Z}}(a))$ *on the interval* $[0, b-a]$.

Proof. Define $\tilde{Q}(\cdot)$, $\tilde{Z}(\cdot)$, $\tilde{B}(\cdot)$ and $\tilde{S}(\cdot,\cdot)$ in the same way as (3.1)–(3.8), then (4.41) and (4.42) imply that

$$\lim_{n\to\infty} \sup_{t\in[0,T]} |\bar{Q}^{r_n}(t) - \tilde{Q}(t)| = 0, \tag{4.44}$$

$$\lim_{n\to\infty} \sup_{t\in[0,T]} |\bar{Z}^{r_n}(t) - \tilde{Z}(t)| = 0, \tag{4.45}$$

$$\lim_{n\to\infty} \sup_{t\in[0,T]} |\bar{B}^{r_n}(t) - \tilde{B}(t)| = 0. \tag{4.46}$$

By (4.43) and (4.45),

$$\sup_{a\le t\le b} \Big| \frac{1}{\bar{Z}^{r_n}(t)} - \frac{1}{\tilde{Z}(t)} \Big| \to 0 \quad \text{as } n \to \infty.$$

Thus, for each $\epsilon > 0$ there exists an $n_0 > 0$ such that

$$\sup_{a\le s<t\le b} |\bar{S}^{r_n}(s,t) - \tilde{S}(s,t)| < \epsilon, \quad \text{for all } n > n_0. \tag{4.47}$$

Since for all r_n, $(\bar{\mathcal{Q}}^{r_n}(\cdot), \bar{\mathcal{Z}}^{r_n}(\cdot))$ satisfies the LPS policy constraints (2.7) and (2.8) and $\frac{K^{r_n}}{r_n} \to K$ as $n \to \infty$, the limit $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ also satisfies (3.12) and (3.13). It is then clear that $(\tilde{\mathcal{Q}}(a), \tilde{\mathcal{Z}}(a))$ is a valid initial condition. By the same argument, $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ also satisfies (3.11). Now, it only remains to show that $(\tilde{\mathcal{Q}}(a+\cdot), \tilde{\mathcal{Z}}(a+\cdot))$ satisfies the fluid dynamic equations (3.9) and (3.10) on the interval $[0, b-a]$.

By (4.16), for any Borel set $A \subset \mathbb{R}_+$ and $t \geq 0$,

$$\bar{\mathcal{Q}}^{r_n}(a+t)(A) = \bar{\mathcal{Q}}^{r_n}(a)(A) + I_0^{r_n}(A) - I_1^{r_n}(A), \tag{4.48}$$

where

$$I_0^{r_n}(A) = \frac{1}{r_n} \sum_{i=r_n \bar{E}^{r_n}(a)+1}^{r_n \bar{E}^{r_n}(a+t)} \delta_{v_i^{r_n}}(A),$$
$$I_1^{r_n}(A) = \frac{1}{r_n} \sum_{i=r_n \bar{B}^{r_n}(a)+1}^{r_n \bar{B}^{r_n}(a+t)} \delta_{v_i^{r_n}}(A).$$

To verify (3.9), consider the following difference for any $y \geq 0$ (recall that $A_y = (y, \infty)$),

$$\begin{aligned}
&\left|\tilde{\mathcal{Q}}(a+t)(A_y) - \left(\tilde{\mathcal{Q}}(a)(A_y) + [\tilde{Q}(a+t) - \tilde{Q}(a)]\nu(A_y)\right)\right| \\
&\leq \left|\tilde{\mathcal{Q}}(a+t)(A_y) - \bar{\mathcal{Q}}^{r_n}(a+t)(A_y)\right| \\
&\quad + \left|\bar{\mathcal{Q}}^{r_n}(a+t)(A_y) - \left(\bar{\mathcal{Q}}^{r_n}(a)(A_y) + I_0^{r_n}(A_y) - I_1^{r_n}(A_y)\right)\right| \\
&\quad + \left|\left(\bar{\mathcal{Q}}^{r_n}(a)(A_y) + I_0^{r_n}(A_y) - I_1^{r_n}(A_y)\right) - \left(\tilde{\mathcal{Q}}(a)(A_y) + [\tilde{Q}(a+t) - \tilde{Q}(a)]\nu(A_y)\right)\right| \\
&\leq \left|\tilde{\mathcal{Q}}(a+t)(A_y) - \bar{\mathcal{Q}}^{r_n}(a+t)(A_y)\right| + \left|\tilde{\mathcal{Q}}(a)(A_y) - \bar{\mathcal{Q}}^{r_n}(a)(A_y)\right| \\
&\quad + \left|[\tilde{Q}(a+t) - \tilde{Q}(a)]\nu(A_y) - I_0^{r_n}(A_y) + I_1^{r_n}(A_y)\right|,
\end{aligned} \tag{4.49}$$

where the first inequality is due to triangle inequality, and the second one is due to (4.48) and another application of triangle inequality. According to Lemma 4.8, the set A_y is a $\tilde{\mathcal{Q}}(a+t)$-continuity set (i.e. a set whose boundary has zero mess under the measure). By Property (iii) on page 72 of [5], the convergence of $\bar{\mathcal{Q}}^{r_n}(a+t)$ to

$\tilde{\mathcal{Q}}(a+t)$ in the Prohorov metric implies weak convergence. By Portmanteau Theorem (c.f. Theorem 2.1 in [5]), weak convergence implies $\bar{\mathcal{Q}}^{r_n}(a+t)(A) \to \tilde{\mathcal{Q}}(a+t)(A)$ for all $\tilde{\mathcal{Q}}(a+t)$-continuity set A. This implies that each of the first two terms on the right hand side of (4.48) can be bounded by ϵ for all large n. Now, let us study the third term. Let $\tilde{E}(\cdot) = \tilde{B}(\cdot) + \tilde{Q}(\cdot)$, so $\tilde{E}(\cdot)$ is the limit of $\bar{E}^r(\cdot)$. (In fact, $\tilde{E}(\cdot) = \lambda\cdot$ as proved in Section 4.1.1. But it is not needed here.) So by triangle inequality, we have that

$$\begin{aligned}
&\left|[\tilde{Q}(a+t) - \tilde{Q}(a)]\nu(A_y) - I_0^{r_n}(A_y) + I_1^{r_n}(A_y)\right| \\
&= \left|[\tilde{E}(a+t) - \tilde{E}(a)]\nu(A_y) - I_0^{r_n}(A_y) - [\tilde{B}(a+t) - \tilde{B}(a)]\nu(A_y) + I_1^{r_n}(A_y)\right| \\
&\le \left|[\tilde{E}(a+t) - \tilde{E}(a)]\nu(A_y) - I_0^{r_n}(A_y)\right| + \left|\tilde{B}(a+t) - \tilde{B}(a)]\nu(A_y) - I_1^{r_n}(A_y)\right|.
\end{aligned}$$

Note that

$$\begin{aligned}
&\left|[\tilde{E}(a+t) - \tilde{E}(a)]\nu(A_y) - I_0^{r_n}(A_y)\right| \\
&\le [\tilde{E}(a+t) - \tilde{E}(a)]\left|\nu(A_y) - \nu^{r_n}(A_y)\right| + \left|[\tilde{E}(a+t) - \tilde{E}(a)]\nu^{r_n}(A_y) - I_0^{r_n}(A_y)\right|.
\end{aligned}$$

Again, since ν is atom free (by condition (4.11)), A_y is a ν-continuity set. So $\left|\nu(A_y) - \nu^{r_n}(A_y)\right| \le \epsilon$ for all large n. Since we restrict our sample path to be in the event $\Omega^{r_n}(M, \mathcal{S})$, and hence in $\Omega_E^{r_n} \cap \Omega_B^{r_n}(M)$ for each n, the limits $\tilde{E}(\cdot)$ and $\tilde{B}(\cdot)$ have an upper bound $M + 2\lambda T$ and a lower bound $-M$ on the interval $[0, T]$. So the first term in the above can be bounded by $(M + 2\lambda T)\epsilon$ for all large n. Note that

$$\begin{aligned}
&\left|[\tilde{E}(a+t) - \tilde{E}(a)]\nu^{r_n}(A_y) - I_0^{r_n}(A_y)\right| \\
&\le \left|\bar{E}^{r_n}(a+t) - \tilde{E}(a+t)\right| + \left|\bar{E}^{r_n}(a) - \tilde{E}(a)\right| \\
&\quad + \left|\frac{1}{r_n}\sum_{i=r_n\tilde{E}(a)+1}^{r_n\tilde{E}(a+t)} \delta_{v_i^{r_n}}(A_y) - [\tilde{E}(a+t) - \tilde{E}(a)]\nu^{r_n}(A_y)\right|
\end{aligned}$$

Since $\tilde{E}(\cdot)$ is the limit of $\bar{E}^{r_n}(\cdot)$, each of the first two terms is bounded by ϵ for all large n. Since we restrict our sample path to be in the event $\Omega^{r_n}(M, \mathcal{S})$, and hence in

$\Omega^{r_n}_{\mathrm{GC}}(M)$ for all n, the last term in the above can be bounded above by ϵ for all large n. Thus, we conclude that

$$\left|[\tilde{E}(a+t)-\tilde{E}(a)]\nu(A_y)-I_0^{r_n}(A_y)\right|\le(M+2\lambda T+3)\epsilon,$$

for all large n. Using exactly the same argument, we can show that

$$\left|[\tilde{B}(a+t)-\tilde{B}(a)]\nu(A_y)-I_1^{r_n}(A_y)\right|\le(M+2\lambda T+3)\epsilon,$$

for all large n. So in summary, the right side of (4.49) is bounded by $(2M+4\lambda T+8)\epsilon$ for all large n. Since $\epsilon>0$ is arbitrary, the left side of (4.49) must be 0. So the fluid dynamic equation (3.9) is verified.

By (4.17), for all Borel set $A\subset\mathbb{R}_+$ and $t\ge 0$,

$$\bar{\mathcal{Z}}^{r_n}(a+t)(A)=\bar{\mathcal{Z}}^{r_n}(a)(A+\bar{S}^{r_n}(a,a+t))+I_2^{r_n}(A), \tag{4.50}$$

where

$$I_2^{r_n}(A)=\frac{1}{r_n}\sum_{i=r_n\bar{B}^{r_n}(a)+1}^{r_n\bar{B}^{r_n}(a+t)}\delta_{v_i^r}(A+\bar{S}^{r_n}(\tau_i^{r_n},a+t)).$$

To verify (3.10), consider the difference

$$\begin{aligned}
&\left|\left(\tilde{\mathcal{Z}}(a+t)(A_y)-\tilde{\mathcal{Z}}(a)(A_y+\tilde{S}(a,a+t))\right)-\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))d\tilde{B}(\tau)\right|\\
\le&\left|\left(\tilde{\mathcal{Z}}(a+t)(A_y)-\tilde{\mathcal{Z}}(a)(A_y+\tilde{S}(a,a+t))\right)\right.\\
&\quad\left.-\left(\bar{\mathcal{Z}}^{r_n}(a+t)(A_y)-\bar{\mathcal{Z}}^{r_n}(a)(A_y+\bar{S}^{r_n}(a,a+t))\right)\right|\\
&+\left|\left(\bar{\mathcal{Z}}^{r_n}(a+t)(A_y)-\bar{\mathcal{Z}}^{r_n}(a)(A_y+\bar{S}^{r_n}(a,a+t))\right)-I_2^{r_n}(A_y)\right|\\
&+\left|\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))d\tilde{B}(\tau)-I_2^{r_n}(A_y)\right|\\
\le&\left|\tilde{\mathcal{Z}}(a+t)(A_y)-\bar{\mathcal{Z}}^{r_n}(a+t)(A_y)\right|\\
&+\left|\tilde{\mathcal{Z}}(a)(A_y+\tilde{S}(a,a+t))-\bar{\mathcal{Z}}^{r_n}(a)(A_y+\bar{S}^{r_n}(a,a+t))\right|\\
&+\left|\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))d\tilde{B}(\tau)-I_2^{r_n}(A_y)\right|,
\end{aligned} \tag{4.51}$$

where the first inequality is due to triangle inequality, and the second one is due to (4.50) and another application of triangle inequality. By Lemma 4.8, the measure $\tilde{\mathcal{Z}}(t+a)$ is also atom free. So following the same argument as the one for $\tilde{\mathcal{Q}}(a)$, the first term on the right hand side in (4.51) is bounded by ϵ for all large n. For any $y \geq 0$ and $\kappa > 0$,

$$\begin{aligned}
&\tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) - \bar{\mathcal{Z}}^{r_n}(a)\big((y+\bar{S}^{r_n}(a,a+t),\infty)\big) \\
&\leq \tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) - \bar{\mathcal{Z}}^{r_n}(a)\big((y+\tilde{S}(a,a+t)+\kappa,\infty)\big) \\
&\leq \tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) - \bar{\mathcal{Z}}^{r_n}(a)\big((y+\tilde{S}(a,a+t)-\kappa,\infty)\big) \\
&\quad + \bar{\mathcal{Z}}^{r_n}(a)\big([y+\tilde{S}(a,a+t)-\kappa, y+\tilde{S}(a,a+t)+\kappa]\big) \\
&\leq \kappa + \bar{\mathcal{Z}}^{r_n}(a)\big([y+\tilde{S}(a,a+t)-\kappa, y+\tilde{S}(a,a+t)+\kappa]\big),
\end{aligned}$$

for all large n, where the first inequality is due to (4.47), the second inequality is due to algebra and the last inequality is due to (4.42) and the definition of Prohorov metric. Since we restrict our sample path to be in the event $\Omega^{r_n}(M,\mathcal{S})$, and hence in $\Omega_R^{r_n}(\mathcal{S})$ for all n, we can choose κ small enough (less than ϵ) to make the second term on the right hand side of the above less than ϵ. Thus we have

$$\tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) - \bar{\mathcal{Z}}^{r_n}(a)\big((y+\bar{S}^{r_n}(a,a+t),\infty)\big) \leq 2\epsilon.$$

On the other side, for any $y \geq 0$ and $\kappa > 0$,

$$\begin{aligned}
&\bar{\mathcal{Z}}^{r_n}(a)\big((y+\bar{S}^{r_n}(a,a+t),\infty)\big) - \tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) \\
&\leq \bar{\mathcal{Z}}^{r_n}(a)\big((y+\tilde{S}(a,a+t)-\kappa,\infty)\big) - \tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) \\
&\leq \bar{\mathcal{Z}}^{r_n}(a)\big([y+\tilde{S}(a,a+t)-\kappa, y+\tilde{S}(a,a+t)+\kappa]\big) \\
&\quad + \bar{\mathcal{Z}}^{r_n}(a)\big((y+\tilde{S}(a,a+t)+\kappa,\infty)\big) - \tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big)
\end{aligned}$$

for all large n, where the first inequality is due to (4.47), the second inequality is due to algebra. By the same argument, we can show that

$$\bar{\mathcal{Z}}^{r_n}(a)\big((y+\bar{S}^{r_n}(a,a+t),\infty)\big) - \tilde{\mathcal{Z}}(a)\big((y+\tilde{S}(a,a+t),\infty)\big) \leq 2\epsilon.$$

This implies that the second term on the right hand side of (4.51) is bounded by 2ϵ. To control the third term, define

$$I_2^{r_n}(A) = \sum_{j=0}^{N-1} I_{2,j}^{r_n}(A),$$

where $0 = t_0 < \ldots < t_{N-1} = t$ is a partition of the interval $[0, t]$ with $\delta = \max_j |t_{j+1} - t_j|$ and

$$I_{2,j}^{r_n}(A) = \frac{1}{r_n} \sum_{i=r_n \bar{B}^{r_n}(a+t_j)+1}^{r_n \bar{B}^{r_n}(a+t_{j+1})} \delta_{v_i^r}(A + \bar{S}^{r_n}(\tau_i^{r_n}, a+t)).$$

Recall that $\tau_i^{r_n}$ is the time that the ith job starts service in the r_nth system, so on each sub-interval $[a+t_j, a+t_{j+1}]$ those i's to be summed must satisfy $a+t_j \le \tau_i^{r_n} \le a+t_{j+1}$. This implies that

$$\bar{S}^{r_n}(a+t_{j+1}, a+t) \le \bar{S}^{r_n}(\tau_i^{r_n}, a+t) \le \bar{S}^{r_n}(a+t_j, a+t).$$

By the uniform convergence (4.47), we have for all large n,

$$y - \epsilon + \tilde{S}(a+t_{j+1}, a+t) \le y + \bar{S}^{r_n}(\tau_i^{r_n}, a+t) \le y + \epsilon + \tilde{S}(a+t_j, a+t).$$

Since we are on the event $\Omega^{r_n}(M, \mathcal{S})$ (which is defined at the end of Section 4.1), for $\epsilon > 0$ there exists an n_1 such that for all $n > n_1$ and $j = 0, \ldots, N-1$,

$$\begin{aligned} I_{2,j}^{r_n}(A_y) &\ge \bar{B}^{r_n}(a+t_j, a+t_{j+1}) \nu^{r_n}(A_y + \epsilon + \tilde{S}(a+t_j, a+t)) - \epsilon, \\ &\ge \tilde{B}(a+t_j, a+t_{j+1}) \nu^{r_n}(A_y + \epsilon + \tilde{S}(a+t_j, a+t)) - 2\epsilon, \\ &\ge \tilde{B}(a+t_j, a+t_{j+1}) \nu(A_y + \tilde{S}(a+t_j, a+t)) - (2M + 2\lambda T + 2)\epsilon, \end{aligned}$$

where the above three inequalities are due to that we are on the event $\Omega_{\mathrm{GC}}^r(M)$, (4.46) and the definition of Prohorov metric, respectively. Please note that the above $2M + 2\lambda T$ comes from $\tilde{B}(s,t) = \tilde{B}(t) - \tilde{B}(s) < 2M + 2\lambda T$, since $\tilde{B}(\cdot)$ has lower bounded $-M$ and upper bound $(M + 2\lambda T)$ (again, because we are on the event $\Omega_B^{r_n}(M)$). By the same reason, for all $n > n_1$ and $j = 0, \ldots, N-1$,

$$I_{2,j}^{r_n}(A_y) \le \tilde{B}(a+t_j, a+t_{j+1}) \nu(A_y + \tilde{S}(a+t_{j+1}, a+t)) + (2M + 2\lambda T + 2)\epsilon.$$

Denoting

$$I_{L,\delta}(A_y) = \sum_{j=0}^{N-1} \tilde{B}(a+t_j, a+t_{j+1})\nu(A_y + \tilde{S}(a+t_j, a+t)),$$
$$I_{U,\delta}(A_y) = \sum_{j=0}^{N-1} \tilde{B}(a+t_j, a+t_{j+1})\nu(A_y + \tilde{S}(a+t_{j+1}, a+t)),$$

we have that

$$I_{L,\delta}(A_y) - N(2M + 2\lambda T + 2)\epsilon \le I_2^{r_n}(A_y) \le I_{U,\delta}(A_y) + N(2M + 2\lambda T + 2)\epsilon. \quad (4.52)$$

It is clear that $I_{L,\delta}(A_y)$ and $I_{U,\delta}(A_y)$ are the Riemann lower sum and upper sum of the integration $\int_a^{a+t} \nu(A_y + \tilde{S}(\tau, a+t))d\tilde{B}(\tau)$, respectively. This means that for all δ small enough,

$$I_{L,\delta}(A_y) \le \int_a^{a+t} \nu(A_y + \tilde{S}(\tau, a+t))d\tilde{B}(\tau) \le I_{U,\delta}(A_y). \quad (4.53)$$

It then follows from (4.52) and (4.53) that

$$\begin{aligned} &\Big| \int_a^{a+t} \nu(A_y + \tilde{S}(\tau, a+t))d\tilde{B}(\tau) - I_2^{r_n}(A_y) \Big| \\ &\le [I_{U,\delta}(A_y) - I_{L,\delta}(A_y)] + 2N(2M + 2\lambda T + 2)\epsilon. \end{aligned} \quad (4.54)$$

For any $\epsilon_1 > 0$, we first choose δ small enough (therefore N is chosen) to make the first term in the upper bound of (4.54) less than $\epsilon_1/2$, and then choose ϵ small enough to make the second term in the upper bound of (4.54) less than $\epsilon_1/2$. So the third term on the right hand side of (4.51) is bounded by ϵ_1. In summary, the right hand side of (4.51) is bounded by $3\epsilon + \epsilon_1$ for all large n. Since $\epsilon, \epsilon_1 > 0$ can be arbitrarily small, the left hand side of (4.51) must be 0. So (3.10) is satisfied. □

Lemma 4.10. *Assume* (4.8)–(4.15). *Fix a point* $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_T(M, \mathcal{S})$ *and a constant* $t_0 \in [0, T]$. *If* $(\tilde{\mathcal{Q}}(t_0), \tilde{\mathcal{Z}}(t_0)) = (\mathbf{0}, \mathbf{0})$, *then*

$$(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0}), \quad \textit{for all } t \in [t_0, T] \quad (4.55)$$

when $\rho \le 1$;

$$\inf_{t \in [t_1, T]} \tilde{Z}(t) > 0, \quad \textit{for all } t_1 \in (t_0, T] \quad (4.56)$$

when $\rho > 1$. *If* $(\tilde{\mathcal{Q}}(t_0), \tilde{\mathcal{Z}}(t_0)) \neq (\mathbf{0}, \mathbf{0})$ *and* $\rho > 1$, *then*

$$\inf_{t\in[t_0,T]} \tilde{Z}(t) > 0. \tag{4.57}$$

Proof. The assumption $(\tilde{\mathcal{Q}}(t_0), \tilde{\mathcal{Z}}(t_0)) = (\mathbf{0}, \mathbf{0})$ implies that

$$\bar{\mathcal{Z}}^{r_n}(t_0) \to \mathbf{0} \text{ as } n \to \infty. \tag{4.58}$$

Note that for any constant $a > 0$, the workload at time t_0 satisfies

$$\begin{aligned}
\bar{W}^{r_n}(t_0) &= \langle \chi, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle \\
&= \langle \chi 1_{[0,a]}, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle + \langle \chi 1_{[a,\infty)}, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle \\
&\leq a\langle 1, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle + \frac{1}{a^p}\langle \chi^{1+p}, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle.
\end{aligned}$$

By the definition of $\Omega_B^{r_n}(M)$, $\langle \chi^{1+p}, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle < M$. For any $\epsilon > 0$, we first choose a large enough such that $\frac{M}{a^p} < \epsilon/2$. By (4.58), we then can choose n large enough such that $a\langle 1, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle \leq \epsilon/2$. This implies that

$$\bar{W}^{r_n}(t_0) = \langle \chi, \bar{\mathcal{Z}}^{r_n}(t_0)\rangle \to 0 \quad \text{as } n \to \infty.$$

By Proposition 4.1,

$$\bar{W}^{r_n}(\cdot) \to \ \bar{W}(\cdot),$$

where $\bar{W}(t) = \big(w^* - (1-\rho)t\big)^+$ for $t \geq 0$. This means that $\bar{W}(t_0) = 0$.

If $\rho \leq 1$, then $\bar{W}(t) = 0$ for all $t \geq t_0$. This means that for each $t \geq t_0$, $\bar{W}^{r_n}(t) \to 0$ as $n \to \infty$. For any $\kappa > 0$, we have the following inequality,

$$\bar{Z}^{r_n}(t) \leq \bar{\mathcal{Z}}^{r_n}(t)(0,\kappa] + \frac{\bar{W}^{r_n}(t)}{\kappa}.$$

Since we are on the event $\Omega_R^{r_n}(\mathcal{S})$ (which is defined at the end of Section 4.1), we can choose κ small enough such that

$$\bar{Z}^{r_n}(t) \leq \epsilon + \frac{\bar{W}^{r_n}(t)}{\kappa},$$

where the second term on the right hand side in the above can be made smaller than ϵ by taking n large enough. This implies that $\tilde{Z}(t) = 0$, which means $(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$.

If $\rho > 1$, then for any $t \in [t_1, T]$ we have

$$\bar{W}(t) \geq (\rho - 1)(t_1 - t_0) \stackrel{\Delta}{=} \alpha_1.$$

Since on the event $\Omega_B^r(M)$ (which is defined in Section 4.1.1), $\langle \chi^{1+p}, \bar{\mathcal{Z}}^{r_n}(t) \rangle < M$ for all $t \in [0, T]$, for any $\epsilon > 0$, there exists a $c_0 > 0$ such that

$$\langle \chi 1_{(c_0, \infty)}, \bar{\mathcal{Z}}^{r_n}(t) \rangle < \epsilon \quad \text{for all } t \in [0, T] \text{ and } n \geq 0.$$

This implies that for all $t \in [0, T]$,

$$\begin{aligned} \bar{W}^{r_n}(t) &= \langle \chi 1_{[0, c_0]}, \bar{\mathcal{Z}}^{r_n}(t) \rangle + \langle \chi 1_{(c_0, \infty)}, \bar{\mathcal{Z}}^{r_n}(t) \rangle \\ &\leq c_0 \bar{Z}^{r_n}(t) + \epsilon. \end{aligned} \tag{4.59}$$

Take $\epsilon = \alpha_1/2$, we have that $\bar{Z}^{r_n}(t) \geq \frac{\alpha_1}{2c_0}$ for all $t \in [t_1, T]$. Letting $n \to \infty$, $\tilde{Z}(t) \geq \frac{\alpha_1}{2c_0}$ for all $t \in [t_1, T]$.

The assumption $(\tilde{\mathcal{Q}}(t_0), \tilde{\mathcal{Z}}(t_0)) \neq (\mathbf{0}, \mathbf{0})$ implies that $\tilde{Z}(t_0) > 0$. If $\rho > 1$, then for any $t \in [t_0, T]$ we have

$$\bar{W}(t) = \bar{W}(t_0) + (\rho - 1)(t - t_0) \geq \bar{W}(t_0) \stackrel{\Delta}{=} \alpha_0 > 0.$$

Note that (4.59) holds on the interval $[0, T]$, we apply it to the interval $[t_0, T]$. Take $\epsilon = \alpha_0/2$, we have that $\bar{Z}^{r_n}(t) \geq \frac{\alpha_0}{2c_0}$ for all $t \in [t_0, T]$. Letting $n \to \infty$, $\tilde{Z}(t) \geq \frac{\alpha_1}{2c_0}$ for all $t \in [t_0, T]$. □

Proof of Theorem 4.3. Case 1, $\rho > 1$. By Lemmas 4.9 and 4.10 for any $0 < t_1 \leq t$, we have that

$$\tilde{\mathcal{Q}}(t)(A_y) = \tilde{Q}(t_1)\nu(A_y) + [\tilde{Q}(t) - \tilde{Q}(t_1)]\nu(A_y), \tag{4.60}$$

$$\tilde{\mathcal{Z}}(t)(A_y) = \tilde{\mathcal{Z}}(t_1)(A_y + \tilde{S}(t_1, t)) + \int_{t_1}^{t} \nu(A_y + \tilde{S}(s, t)) d\bar{B}(s), \tag{4.61}$$

for all $y \geq 0$. Since $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is continuous, we have that

$$(\tilde{\mathcal{Q}}(t_1), \tilde{\mathcal{Z}}(t_1)) \to (\xi^*, \mu^*) \quad \text{as } t_1 \to 0. \tag{4.62}$$

So letting $t_1 \to 0$, (4.60) becomes

$$\tilde{\mathcal{Q}}(t)(A_y) = \xi^*(A_y) + [\tilde{Q}(t) - \tilde{Q}(0)]\nu(A_y).$$

Note that

$$\begin{aligned}&\int_0^{t_1} \nu(A_y + \tilde{S}(s,t))d\bar{B}(s)\\ &\leq [\tilde{B}(t_1) - \tilde{B}(0)] = \lambda t_1 - (\tilde{Q}(t_1) - \tilde{Q}(0)),\end{aligned}$$

which converges to 0 as $t_1 \to 0$. If $\mu^* \neq \mathbf{0}$, then $\tilde{Z}(0) > 0$. Lemma 4.10 implies that $\inf_{s\in[0,t]} \tilde{Z}(s) > 0$. This implies that

$$\tilde{S}(t_1, t) \to \tilde{S}(t) \quad \text{as } t_1 \to 0. \tag{4.63}$$

By (4.62), $\tilde{\mathcal{Z}}(t_1) \to \mu^*$ (in the Prohorov metric) as $t_1 \to 0$. It follows from (4.63) that

$$\tilde{\mathcal{Z}}(t_1)(A_y + \tilde{S}(t_1, t)) \to \mu^*(A_y + \tilde{S}(t)) \quad \text{as } t_1 \to 0.$$

If $\mu^* = \mathbf{0}$,

$$\tilde{\mathcal{Z}}(t_1)(A_y + \tilde{S}(t_1, t)) \to 0 \quad \text{as } t_1 \to 0.$$

In both cases, letting $t_1 \to 0$, (4.61) becomes

$$\tilde{\mathcal{Z}}(t)(A_y) = \mu^*(A_y + \tilde{S}(t)) + \int_0^t \nu(A_y + \tilde{S}(s,t))d\bar{B}(s).$$

So we conclude that $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is the fluid model solution with initial condition (ξ^*, μ^*).

Case 2, $\rho \leq 1$. Let $t_0 = \inf\{t \geq 0 : \tilde{\mathcal{Z}}(t) = \mathbf{0}\}$. By Lemma 4.9, for all $t \in [0, t_0)$, $(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t))$ satisfies the fluid dynamic equations (3.9) and (3.10) with initial condition (ξ^*, μ^*). By the continuity of $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ and Lemma 4.10, $(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$ for all $t \in [t_0, T]$. So $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is the fluid model solution with initial condition (ξ^*, μ^*). □

Proof of Theorem 4.1. It is enough to show that for any $\eta, \epsilon > 0$,

$$\liminf_{r\to\infty} \mathbb{P}\left(\varrho[(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot)), (\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))] < \epsilon\right) \geq 1 - \eta,$$

where ϱ is the Skorohod metric defined in Section 1.4. Fix a constant $T > 0$ such that $\int_T^\infty e^{-t}dt < \epsilon/2$. By Theorem 5.3 and Theorem 4.3, we have that

$$\liminf_{r\to\infty} \mathbb{P}\left(\varrho_T[(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot)), (\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))] < \frac{\epsilon}{2(1 - e^{-T})}\right) \geq 1 - \eta.$$

The result follows immediately from (1.4). □

CHAPTER V

STATE SPACE COLLAPSE AND DIFFUSION LIMITS

In this Chapter, we study a "finer" scaling of the LPS queueing system, called the *diffusion* scaling. It is defined by

$$\hat{\mathcal{Q}}^r(t) = \frac{1}{r}\mathcal{Q}^r(r^2t), \quad \hat{\mathcal{Z}}^r(t) = \frac{1}{r}\mathcal{Z}^r(r^2t), \tag{5.1}$$

for all $t \geq 0$. As it has been shown in Williams [57], a key step to obtain a diffusion limit in heavy traffic is to establish a *state space collapse* (SSC) result. In our setting, the SSC means that the diffusion-scaled measure-valued process, which is an infinite dimensional object, is close to a deterministic functional of the diffusion-scaled, one-dimensional workload process. (See Definition 3.4 for the lifting map to define the functional.) Our proof strategy is analogous to the modular approach proposed in Bramson [7] and Williams [57].

We are also interested in other diffusion scaled quantities like the workload and queue length processes. Note that $Q^r(\cdot)$, $Z^r(\cdot)$ and $W^r(\cdot)$ are actually functions of $(\mathcal{Q}^r(\cdot), \mathcal{Z}^r(\cdot))$, so the scaling for these quantities is defined as the functions of the corresponding scaling for $(\mathcal{Q}^r(\cdot), \mathcal{Z}^r(\cdot))$, i.e.

$$\hat{Q}^r(t) = \langle 1, \hat{\mathcal{Q}}^r(t)\rangle = \frac{1}{r}Q^r(r^2t), \tag{5.2}$$

$$\hat{Z}^r(t) = \langle 1, \hat{\mathcal{Z}}^r(t)\rangle = \frac{1}{r}Z^r(r^2t), \tag{5.3}$$

$$\hat{W}^r(t) = \langle \chi, \hat{\mathcal{Q}}^r(t) + \hat{\mathcal{Z}}^r(t)\rangle = \frac{1}{r}W^r(r^2t), \tag{5.4}$$

for all $t \geq 0$.

To establish results on the convergence of the above sequence of stochastic processes, we need the following conditions in addition to conditions (4.9)–(4.15), which

are quite general and standard. We assume that the arrival processes satisfy

$$\frac{E^r(r^2\cdot) - \lambda^r r^2 \cdot}{r} \Rightarrow E^*(\cdot) \quad \text{as } r \to \infty, \tag{5.5}$$

where

$$\lim_{r\to\infty} \lambda^r = \lambda > 0, \tag{5.6}$$

and $E^*(\cdot)$ is a Brownian motion with drift 0 and variance λc_a^2. And the measures of job sizes satisfy that as $r \to \infty$.

$$\langle \chi^{2+2p}, \nu^r \rangle \to \langle \chi^{2+2p}, \nu \rangle \quad \text{for some } p > 0. \tag{5.7}$$

Define the traffic intensity of the rth system by $\rho^r = \lambda^r \langle \chi, \nu^r \rangle$. We need the following heavy traffic condition:

$$\lim_{r\to\infty} r(1 - \rho^r) = \theta > 0. \tag{5.8}$$

Let $\beta = \langle \chi, \nu \rangle$ be the mean and $c_s^2 = \frac{\langle \chi^2, \nu \rangle - \beta^2}{\beta^2}$ be the squared coefficient of variation (SCV) of the job size distribution ν. The following proposition is a well-known heavy traffic approximation for the workload process of a single queue operated under a non-idling policy. Readers are referred to [23] for a proof.

Proposition 5.1. *Assume* (4.9)–(4.15), (5.5) *and* (5.8). *The sequence of diffusion scaled workload process*

$$\hat{W}^r(\cdot) \Rightarrow W^*(\cdot) \quad \textit{as } r \to \infty,$$

where $W^*(\cdot)$ is a reflected Brownian motion with drift $-\theta$, variance $\beta(c_a^2 + c_s^2)$ and initial value $w^* = \langle \chi, \xi^* + \mu^* \rangle$.

Since the LPS is also a non-idling service policy, the above result on the workload process is still true for our model. However, the diffusion limit for job size process $X(\cdot)$ and many other performance processes as introduced in Chapter 2 do not follow from this result as the queue length process does depend on the service discipline. Our main

result establishes the diffusion limit for the measure valued processes (Theorem 5.1), from which the diffusion limit of queue length process follows directly (Corollary 5.1).

Our main results require that the limit (ξ^*, μ^*) in (4.13) satisfies

$$(\xi^*, \mu^*) = \Delta_{K,\nu} w^*, \tag{5.9}$$

where $\Delta_{K,\nu}$ is defined in Definition 3.4.

Theorem 5.1. *Assume* (4.9)–(4.15), (5.5)–(5.9). *The sequence of diffusion scaled state descriptors*

$$(\hat{\mathcal{Q}}^r(\cdot), \hat{\mathcal{Z}}^r(\cdot)) \Rightarrow \Delta_{K,\nu} W^*(\cdot) \quad \textit{as } r \to \infty,$$

where $W^*(\cdot)$ is the reflected Brownian motion in Proposition 3.1.

The major step in establishing the diffusion limit is the following *state space collapse* result.

Theorem 5.2. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix $T > 0$, we have*

$$\sup_{t\in[0,T]} \mathbf{d}[(\hat{\mathcal{Q}}^r(t), \hat{\mathcal{Z}}^r(t)), \Delta_{K,\nu} \hat{W}^r(t)] \Rightarrow 0 \quad \textit{as } r \to \infty.$$

The state-space collapse result is appealing, since it rigorously shows that all performance processes can be described as a simple, deterministic functional of the workload process. We now use Theorem 5.2 to prove the main result.

Proof of Theorem 5.1. We have the convergence of the workload processes $\hat{W}^r(t)$ in Proposition 3.1. Since the mapping $\Delta_{K,\nu} : \mathbb{R}_+ \to \mathbf{M}_1 \times \mathbf{M}_2$ is continuous, by the continuous mapping theorem

$$\Delta_{K,\nu} \hat{W}^r(\cdot) \Rightarrow \Delta_{K,\nu} W^*(\cdot) \quad \text{as } r \to \infty.$$

The result of the theorem follows immediately from the state space collapse result in Theorem 5.2 and the "convergence together lemma" (Theorem 4.1 in [5]). □

Corollary 5.1 (Piecewise Reflected Brownian Motion)**.** *Assume* (4.9)–(4.15), (5.5)–(5.9)*. The sequence of diffusion scaled total job size process* $\hat{X}^r(\cdot) = \langle 1, \hat{\mathcal{Q}}^r(\cdot) + \hat{\mathcal{Z}}^r(\cdot)\rangle$ *converges in distribution as* $r \to \infty$ *to* $X^*(\cdot)$*, where*

$$X^*(t) = \frac{(W^*(t) - K\beta_e)^+}{\beta} + \frac{W^*(t) \wedge K\beta_e}{\beta_e} \quad \textit{for } t \geq 0,$$

and $W^*(\cdot)$ *is the reflected Brownian motion as in Proposition 3.1.*

Proof. Since $\hat{X}^r(\cdot) = \langle 1, \hat{\mathcal{Q}}^r(\cdot) + \hat{\mathcal{Z}}^r(\cdot)\rangle$ and the mapping $\Phi : \mathbf{M} \to \mathbb{R}$ defined by $\Phi(\cdot) = \langle 1, \cdot\rangle$ is continuous, the result follows from Theorem 5.1 and the continuous mapping theorem. □

Remark 5.1. *In other words,* $X^*(\cdot)$ *is a reflected Brownian motion with drift* $\frac{-\theta}{\beta}$ *and variance* $\frac{c_a^2+c_s^2}{\beta^2}$ *when it is above* $K\beta_e$ *and with drift* $\frac{-\theta}{\beta_e}$ *and variance* $\frac{c_a^2+c_s^2}{\beta_e^2}$ *when it is below* $K\beta_e$*.*

5.1 Shifted Fluid Scaling and Precompactness

5.1.1 Shifted Fluid Scaling

Much of our understanding of the diffusion scaled process will be derived from results about the *shifted fluid scaled process*, which is defined by

$$\bar{\mathcal{Q}}^{r,m}(t) = \frac{1}{r}\mathcal{Q}^r(rm + rt), \quad \bar{\mathcal{Z}}^{r,m}(t) = \frac{1}{r}\mathcal{Z}^r(rm + rt), \tag{5.10}$$

for all $m \in \mathbb{N}$ and $t \geq 0$. To see the relationship between these two scalings, consider the diffusion scaled process on the interval $[0, T]$, which corresponds to the interval $[0, r^2T]$ for the unscaled process. Fix a constant $L > 1$, the interval will be covered by the $\lfloor rT \rfloor + 1$ overlapping intervals

$$[rm, rm + rL] \quad m = 0, 1, \cdots, \lfloor rT \rfloor.$$

For each $t \in [0, T]$, there exists an $m \in \{0, \cdots, \lfloor rT \rfloor\}$ and $s \in [0, L]$ (which may not be unique) such that $r^2t = rm + rs$. Thus

$$\hat{\mathcal{Q}}^r(t) = \bar{\mathcal{Q}}^{r,m}(s), \quad \hat{\mathcal{Z}}^r(t) = \bar{\mathcal{Z}}^{r,m}(s). \tag{5.11}$$

This will serve as a key relationship between fluid and diffusion scaled processes. The above is the idea in the framework of Bramson [7] on how to translate the fluid model SSC result into the diffusion-scaled SSC result. In this chapter, we are following this idea.

We are also interested in shifted fluid scaled versions of other processes, like the workload and job size processes. Note that $Q^r(\cdot)$, $Z^r(\cdot)$, $X^r(\cdot)$, $W^r(\cdot)$ and $S^r(\cdot,\cdot)$ are actually functions of $(\mathcal{Q}^r(\cdot), \mathcal{Z}^r(\cdot))$, so the scaling for these quantities is defined as the functions of the corresponding scaling for $(\mathcal{Q}^r(\cdot), \mathcal{Z}^r(\cdot))$, i.e.

$$\bar{Q}^{r,m}(t) = \langle 1, \bar{\mathcal{Q}}^{r,m}(t)\rangle = \frac{1}{r}Q^r(rm+rt), \tag{5.12}$$

$$\bar{Z}^{r,m}(t) = \langle 1, \bar{\mathcal{Z}}^{r,m}(t)\rangle = \frac{1}{r}Z^r(rm+rt), \tag{5.13}$$

$$\bar{X}^{r,m}(t) = \langle 1, \bar{\mathcal{Z}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle = \frac{1}{r}X^r(rm+rt), \tag{5.14}$$

$$\bar{W}^{r,m}(t) = \langle \chi, \bar{\mathcal{Q}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle = \frac{1}{r}W^r(rm+rt), \tag{5.15}$$

$$\bar{S}^{r,m}(s,t) = \int_s^t \psi(\bar{Z}^{r,m}(\tau))d\tau = \int_{rm+rs}^{rm+rt} \psi(Z^r(\tau))d\tau, \tag{5.16}$$

for all $0 \le s \le t$. We define the shifted fluid scaling for the arrival process as

$$\bar{E}^{r,m}(t) = \frac{1}{r}E^r(rm+rt). \tag{5.17}$$

for all $t \ge 0$. By (2.3), the shifted fluid scaling for $B^r(\cdot)$ is

$$\bar{B}^{r,m}(t) = \bar{E}^{r,m}(t) - \bar{Q}^{r,m}(t), \tag{5.18}$$

for all $t \ge 0$. To shorten the notation, for all $0 \le s \le t$, denote

$$\bar{E}^{r,m}(s,t) = \bar{E}^{r,m}(t) - \bar{E}^{r,m}(s), \quad \bar{B}^{r,m}(s,t) = \bar{B}^{r,m}(t) - \bar{B}^{r,m}(s). \tag{5.19}$$

A shifted fluid scaled version of the stochastic dynamic equations (2.5) and (2.6) can be written as

$$\bar{\mathcal{Q}}^{r,m}(t)(A) = \bar{\mathcal{Q}}^{r,m}(s)(A) + \frac{1}{r}\sum_{i=r\bar{E}^{r,m}(s)+1}^{r\bar{E}^{r,m}(t)} \delta_{v_i^r}(A) - \frac{1}{r}\sum_{i=r\bar{B}^{r,m}(s)+1}^{r\bar{B}^{r,m}(t)} \delta_{v_i^r}(A), \tag{5.20}$$

$$\bar{\mathcal{Z}}^{r,m}(t)(A) = \bar{\mathcal{Z}}^{r,m}(s)(A + \bar{S}^{r,m}(s,t)) + \frac{1}{r}\sum_{i=r\bar{B}^{r,m}(s)+1}^{r\bar{B}^{r,m}(t)} \delta_{v_i^r}(A + \bar{S}^{r,m}(\tau_i^r, t)), \tag{5.21}$$

for $0 \le s \le t$ and a Borel set A. The dynamics of the system is determined by the above equations. Equation (5.20) says that the status of the buffer at time t equals the status at time s plus what has arrived to the buffer and minus what has left from the buffer during time interval $(s, t]$. Those jobs who left buffer enter service; the service process has been taken care of by shifting the set A by the cumulative service amount $\bar{S}^{r,m}(\tau_i, t)$ that the ith job receives. This corresponds to the second term on the right hand side of (5.21). This plus the status at time s shifted by accumulative service amount $\bar{S}^{r,m}(s, t)$ is equal to the status of the server at time t, as indicated in (5.21).

5.1.2 Preliminary Estimates

We first establish some bounds which will be useful for later discussion. The following lemma gives some bound on the arrival processes.

Lemma 5.1. *Assume* (5.5) *and* (5.6). *Fix* $T > 0$ *and* $L > 1$. *For all* $\epsilon, \epsilon' > 0$, *there exists an* r_0 *such that whenever* $r \ge r_0$,

$$\mathbb{P}\left(\max_{m \le \lfloor rT \rfloor} \sup_{s,t \in [0,L]} |E^{r,m}(s,t) - \lambda(t-s)| > \epsilon' \right) < \epsilon. \tag{5.22}$$

Proof. Let $t' = \frac{m+t}{r}$ and $s' = \frac{m+s}{r}$. Note that $\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} \frac{m+t}{r} < T + 1$ for all large r, and $0 \le s, t \le L$ is the same as $|t' - s'| \le L/r$. For any $\delta > 0$, there exists an r_0' such that $L/r < \delta$ for all $r \ge r_0$, so the left hand of (5.22) can be bounded above by

$$\mathbb{P}\left(\sup_{s',t' \in [0,T+1], |s'-t'| < \delta} |\frac{1}{r} E^r(r^2 t') - \lambda r t' - (\frac{1}{r} E^r(r^2 s') - \lambda r s')| > \epsilon' \right) \tag{5.23}$$

for all $r \ge r_0'$. By the assumptions (5.5) and (5.6) on the arrival process, $\{\frac{1}{r} E(r^2 \cdot) - \lambda r \cdot\}$ converges in distribution to the Brownian motion $E^*(\cdot)$. Since a Brownian motion is almost surely continuous, we conclude that (5.23) converges to zero as $\delta \to 0$. Then the inequality (5.22) follows immediately. □

By the same reason as in Lemma 4.1, the ϵ' and ϵ in (5.22) can be replaced by $\epsilon_E(r)$, which is a function of r that vanishes at infinity. Based on this, we construct

$$\Omega_E^r = \{ \max_{m \le \lfloor rT \rfloor} \sup_{s,t \in [0,L]} |E^{r,m}(s,t) - \lambda(t-s)| < \epsilon_E(r) \}. \tag{5.24}$$

According to Lemma 5.1, we have that

$$\lim_{r \to \infty} \mathbb{P}\left(\Omega_E^r\right) = 1. \tag{5.25}$$

Recall the Glivenko-Cantelli estimate in Lemma D.2. By the same argument as in the above, for fixed constant M_1, L_1, there exists a function $\epsilon_{\mathrm{GC}}(\cdot)$, which vanishes at infinity, such that the probability inequality in Lemma D.2 holds with ϵ and ϵ' replaced by this function. In other words, if we denote

$$\Omega_{\mathrm{GC}}^r(M_1, L_1) = \{ \max_{-rM_1 < n < r^2 M_1} \sup_{l \in [0, L_1]} \sup_{f \in \mathscr{V}} |\langle f, \bar{\eta}^r(n,l) \rangle - l \langle f, \nu^r \rangle| < \epsilon_{\mathrm{GC}}(r) \}, \tag{5.26}$$

where $\bar{\eta}^r(n,l)$ is defined in (D.1) and $\mathscr{V}$ is defined in Appendix D, then for any fixed constant M_1, L_1,

$$\lim_{r \to \infty} \mathbb{P}\left(\Omega_{\mathrm{GC}}^r(M_1, L_1)\right) = 1. \tag{5.27}$$

Now we use the above result and Proposition 3.1 to obtain a bound on the queue length processes.

Lemma 5.2. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $T > 0$ *and* $L > 1$. *For all* $\eta > 0$ *there exists a constant* $M > 0$ *such that*

$$\liminf_{r \to \infty} \mathbb{P}\left(\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} \bar{Q}^{r,m}(t) < M \right) \ge 1 - \eta.$$

Proof. Since $\frac{\lfloor rT \rfloor + L}{r} < T + 1$ for all large enough r, it is enough to prove the following inequality:

$$\liminf_{r \to \infty} \mathbb{P}\left(\sup_{t \in [0, T+1]} \hat{Q}^r(t) < M \right) \ge 1 - \eta.$$

Suppose this is not true, then for any M there exists a $\eta > 0$, such that

$$\liminf_{r \to \infty} \mathbb{P}\left(\sup_{t \in [0, T+1]} \hat{Q}^r(t) > M \right) > \eta.$$

Denote the event in the above probability by Ω_1^r. By the stochastic dynamic equation (2.5), we have

$$\frac{1}{r}\mathcal{Q}(r^2t)(A) = \frac{1}{r}\sum_{i=B^r(r^2t)+1}^{E^r(r^2t)} \delta_{v_i^r}(A).$$

Since ν is a probability measure on $\mathbb{R}_+$, there exists an $a > 0$ such that $\nu(a,\infty) > 0$. We have the following inequality from the dynamic equation (5.20),

$$\frac{1}{r}W^r(r^2t) > a\frac{1}{r}\mathcal{Q}^r(r^2t)(a,\infty) \geq \frac{a}{r}\sum_{i=B^r(r^2t)+1}^{E^r(r^2t)} \delta_{v_i^r}(a,\infty). \tag{5.28}$$

For any r, on the event Ω_1^r there exists a $t_1 \in [0, T+1]$ (random and depending on r) such that

$$\frac{1}{r}Q^r(r^2t_1) > M.$$

By (5.24), on the event Ω_E^r,

$$\sup_{t\in[0,T+1]} E^r(r^2t) \leq 2\lambda r^2(T+1),$$

for all large enough r. Let $M_1 = \max(M, 2\lambda T)$ and $L_1 = M$. By (5.26) and (5.28), on the event $\Omega_{\mathrm{GC}}^r(M_1, L_1) \cap \Omega_E^r \cap \Omega_1^r$,

$$\hat{W}^r(t_1) > aM\nu^r(a,\infty) > aM\nu(a,\infty)/2,$$

for all large r. By (5.25) and (5.27), we have

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T+1]} \hat{W}^r(t) > aM\nu(a,\infty)/2\right) > \eta.$$

This contradicts the result in Proposition 3.1. □

The following lemma gives a bound on the $(1+p)$th moment of the measure valued process, where p is the same as in conditions (4.10) and (4.14).

Lemma 5.3. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $T > 0$ *and* $L > 1$. *For each* $\eta > 0$ *there exists a constant* $M > 0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\max_{m\leq\lfloor rT\rfloor} \sup_{t\in[0,L]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle < M\right) \geq 1-\eta.$$

Proof. By condition (4.14),

$$\liminf_{r\to\infty} \mathbb{P}\left(\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(0)\rangle < \langle \chi^{1+p}, \mu^*\rangle + 1\right) = 1.$$

Denote the event in the above by Ω_0^r. By Lemma 5.2, for any $\eta > 0$, there exists a constant $M' > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \frac{1}{r} Q^r(rm+rt) < M'\right) > 1 - \eta/2.$$

Denote the event in the above by $\Omega_1^r(M)$. Fix $M_1 = \max(M', \lambda(T+1))$ and $L_1 = \lambda(L+1) + 2M'$. By Lemma D.2,

$$\lim_{r\to\infty} \mathbb{P}\left(\Omega^r_{\mathrm{GC}}(M_1, L_1)\right) = 1.$$

To prove the lemma, it suffices to show that there exists an $M > 0$ such that on the event $\Omega_0^r \cap \Omega_1^r(M') \cap \Omega^r_{\mathrm{GC}}(M_1, L_1) \cap \Omega^r_E$,

$$\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt) + \frac{1}{r}\mathcal{Z}^r(rm+rt)\rangle < M,$$

for all large r. In the remainder of the proof, all random quantities of the rth system is evaluated at a sample path in the event $\Omega_0^r \cap \Omega_1^r(M') \cap \Omega^r_{\mathrm{GC}}(M_1, L_1) \cap \Omega^r_E$.

We first find a bound for $\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt)\rangle$. By the dynamic equation (2.5), we have that for all $m \le \lfloor rT\rfloor$ and $t \in [0, L]$,

$$\langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt)\rangle = \langle \chi^{1+p}, \frac{1}{r}\sum_{B^r(rm+rt)}^{E^r(rm+rt)} \delta_{v_i^r}\rangle.$$

By (5.24) and the definition of $\Omega_1^r(M')$, we have

$$\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} E^r(rm+rt) < \lambda r^2(T+1) \le r^2 M_1, \tag{5.29}$$

$$\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} Q^r(rm+rt) < rM' \le rL_1. \tag{5.30}$$

for all large enough r. So

$$\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt)\rangle \le \sup_{-rM_1<n<r^2M_1} \langle \chi^{1+p}, \bar{\eta}^r(n, L_1)\rangle.$$

By the remark after Lemma D.2, the function $\chi^{1+p} \in \mathscr{V}$, which appears in the definition of $\Omega^r(M_1, L_1)$. So by (5.26) and (4.10), for all large r,

$$\begin{aligned}\max_{m\le\lfloor rT\rfloor}\sup_{t\in[0,L]}\langle\chi^{1+p},\bar{\mathcal{Q}}^{r,m}(t)\rangle &\le L_1\langle\chi^{1+p},\nu^r\rangle+1/2\\ &\le L_1\langle\chi^{1+p},\nu\rangle+1.\end{aligned}$$

We now look for a bound for $\max_{m\le\lfloor rT\rfloor}\sup_{t\in[0,L]}\langle\chi^{1+p},\frac{1}{r}\mathcal{Z}^r(rm+rt)\rangle$. It follows from the dynamic equation (2.6) that for any $m \le \lfloor rT\rfloor$, $t\in[0,L]$ and Borel set $A\subset\mathbb{R}_+$,

$$\begin{aligned}\frac{1}{r}Z^r(rm+rt)(A) &= \frac{1}{r}Z^r(0)(A+S^r(0,rm+rt))\\ &\quad+\sum_{j=0}^{m-1}\frac{1}{r}\sum_{i=B^r(r(m-j-1))+1}^{B^r(r(m-j))}\delta_{v_i^r}(A+S^r(\tau_i^r,rm+rt))\\ &\quad+\frac{1}{r}\sum_{i=\bar{B}^r(rm)+1}^{B^r(rm+rt)}\delta_{v_i^r}(A+S^r(\tau_i^r,rm+rt)).\end{aligned}$$

Given $0\le j\le m-1$, for those i's with $B^r(r(m-j-1))<i\le B^r(r(m-j))$ we have

$$\tau_i^r\in[r(m-j-1),r(m-j)].$$

Thus, by (5.16), the cumulative service amount that the ith job receives by time $rm+t$ satisfies

$$S^r(\tau_i^r,rm+rt)\ge S^r(r(m-j),rm)\ge\frac{rj}{K^r}\ge\frac{j}{2K},$$

where the last inequality is due to (4.12). For those i's such that τ_i^r larger than $\bar{B}^r(rm)$, we use the trivial lower bound $S^r(\tau_i^r,rm+rt)\ge 0$. Also take the trivial lower bound that $S^r(0,rm+rt)\ge 0$. Then we have the following inequality on the

$(1+p)$th moment

$$\begin{aligned}\langle \chi^{1+p}, \frac{1}{r} Z^r(rm+rt)\rangle \le \langle \chi^{1+p}, \frac{1}{r} Z^r(0)\rangle \\ + \sum_{j=0}^{m-1} \langle \big((\chi - \frac{j}{2K})^+\big)^{1+p}, \frac{1}{r} \sum_{i=B^r(r(m-j-1))+1}^{B^r(r(m-j))} \delta_{v_i^r}\rangle \\ + \langle \chi^{1+p}, \frac{1}{r} \sum_{i=B^r(rm)+1}^{B^r(rm+rt)} \delta_{v_i^r}\rangle.\end{aligned}$$

By (5.29) and (5.30), for all $m \le \lfloor rT \rfloor$, $t \in [0, L]$ and all large r,

$$\begin{aligned}&-rM' \le B^r(rj) \le \lambda r^2(T+1) \le r^2 M_1 \\ &0 \le B^r(rj+rt) - B^r(rj) \le \lambda r(L+1) + 2rM' < rL_1 \quad \text{for all } t \in [0, L].\end{aligned}$$

So

$$\begin{aligned}\langle \chi^{1+p}, \frac{1}{r} \mathcal{Z}^r(rm+rt)\rangle \le \langle \chi^{1+p}, \frac{1}{r} \mathcal{Z}^r(0)\rangle \\ + \sup_{-rM_1 < n < r^2 M_1} \sum_{j=0}^{m-1} \langle \big((\chi - \frac{j}{2K})^+\big)^{1+p}, \bar{\eta}^r(n, L_1)\rangle \qquad (5.31) \\ + \sup_{-rM_1 < n < r^2 M_1} \langle \chi^{1+p}, \bar{\eta}^r(n, L_1)\rangle.\end{aligned}$$

The first term in the above is bounded by $\langle \chi^{1+p}, \mu^*\rangle + 1$ by the definition of Ω_0^r. The third term in the above is bounded by

$$L_1 \langle \chi^{1+p}, \nu^r\rangle + 1/2 \le L_1 \langle \chi^{1+p}, \nu\rangle + 1, \qquad (5.32)$$

for all large r by (5.26) and (4.10). It now only remains to deal with the second term in (5.31). Let

$$\bar{F}_n^r(x) = \bar{\eta}^r(n, L_1)((x, \infty)) \quad \text{for all } x \ge 0.$$

The summation in the second term in (5.31) can be upper bounded by

$$\begin{aligned}&\frac{1}{1+p} \sum_{j=1}^{m-1} \int_{\frac{j}{2K}}^{\infty} (x - \frac{j}{2K})^p \bar{F}_n^r(x) dx \\ &\le \frac{2K}{1+p} \sum_{j=1}^{m-1} \int_{\frac{j-1}{2K}}^{\frac{j}{K}} \int_y^{\infty} (x-y)^p \bar{F}_n^r(x) dx dy \\ &\le \frac{1}{1+p} \int_0^{\infty} \int_y^{\infty} (x-y)^p \bar{F}_n^r(x) dx dy.\end{aligned}$$

Applying Fubini's theorem, the above can be further bounded by

$$\begin{aligned}&\frac{1}{1+p}\int_0^\infty\int_0^x (x-y)^p \bar{F}_n^r(x)dydx\\ \le\ &\frac{1}{(1+p)^2}\int_0^\infty x^{1+p}\bar{F}_n^r(x)dx = \frac{2+p}{(1+p)^2}\langle \chi^{2+p}, \bar{\eta}^r(n, L_1)\rangle.\end{aligned}$$

It again follows from (5.26) and (4.10) that the second term in (5.31) is bounded by

$$\frac{2+p}{(1+p)^2}L_1\langle \chi^{1+p}, \nu^r\rangle + 1/2 \le \frac{2+p}{(1+p)^2}L_1\langle \chi^{1+p}, \nu\rangle + 1,$$

for all large r. The proof of this lemma is completed by summing up all these upper bounds. □

For any r, the rth system is defined on the probability space $(\Omega^r, \mathbb{P}^r, \mathcal{F}^r)$. The stochastic processes $\mathcal{Q}^r(\cdot)$ and $\mathcal{Z}^r(\cdot)$ are actually measurable functions on Ω^r. From now on, we explicitly write these processes down in the form of $\mathcal{Q}^r(\omega, \cdot)$ and $\mathcal{Z}^r(\omega, t)$ to indicate that they are evaluated on the sample path $\omega \in \Omega^r$.

The following proposition summarizes the bound estimates in this section.

Proposition 5.2. *Assume* (4.9)–(4.15), (5.5)–(5.9). *For any $\eta > 0$, there exists a constant $M > 0$ and an event $\Omega_B^r(M)$ for each index r such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_B^r(M)\right) > 1 - \eta, \tag{5.33}$$

and for all $\omega \in \Omega_B^r(M)$ and $r \in \mathbb{R}_+$,

$$\begin{aligned}&\max_{m\le\lfloor rT\rfloor}\sup_{t\in[0,L]} \bar{Q}^{r,m}(\omega, t) < M,\\ &\max_{m\le\lfloor rT\rfloor}\sup_{t\in[0,L]} \bar{W}^{r,m}(\omega, t) < M,\\ &\max_{m\le\lfloor rT\rfloor}\sup_{t\in[0,L]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(\omega, t) + \bar{\mathcal{Z}}^{r,m}(\omega, t)\rangle < M.\end{aligned}$$

Proof. The first and the third inequality follow from Lemmas 5.2 and 5.3. The second inequality follows from Proposition 3.1. □

5.1.3 Compact Containment

A set $\mathbf{K} \subset \mathbf{M}$ is relatively compact if

$$\sup_{\xi \in \mathbf{K}} \xi(\mathbb{R}_+) < \infty,$$

and there exists a sequence of nested compact sets $J_n \subset \mathbb{R}_+$ such that $\bigcup J_n = \mathbb{R}_+$ and

$$\lim_{n \to \infty} \sup_{\xi \in \mathbf{K}} \xi(J_n^c) = 0,$$

where J_n^c denotes the complement of J_n; see [33], Theorem A7.5. We establish the following relative compactness property using the bound estimates in Section 5.1.2.

Lemma 5.4. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $T > 0$ *and* $L > 1$. *For each* $\eta > 0$ *there exist a constant* $M > 0$ *and a compact set* $\mathbf{K}(M) \subset \mathbf{M}$ *such that for all* $\omega \in \Omega_B^r(M)$ *(which is introduced in Proposition 5.2) and* $r \in \mathbb{R}_+$,

$$\bar{\mathcal{Q}}^{r,m}(\omega, t) \in \mathbf{K}(M) \text{ and } \bar{\mathcal{Z}}^{r,m}(\omega, t) \in \mathbf{K}(M) \text{ for all } t \in [0, T] \text{ and } m \leq \lfloor rT \rfloor.$$

Proof. Let

$$\mathbf{K}(M) = \Big\{ \xi \in \mathbf{M} : \xi(\mathbb{R}_+) < M \text{ and } \xi((n, \infty)) \leq M/n^{1+p} \Big\}.$$

Clearly, $\mathbf{K}(M)$ is a compact set for any constant $M > 0$. Note that $\bar{\mathcal{Q}}^{r,m}(\omega, t)(\mathbb{R}_+)$ is bounded by M for all $m \leq \lfloor rT \rfloor$, $t \in [0, T]$ and $\omega \in \Omega_B^r(M)$. By the Markov inequality, for any $t \geq 0$, $m \leq \lfloor rT \rfloor$ and $\omega \in \Omega_B^r(M)$,

$$\bar{\mathcal{Q}}^{r,m}(\omega, t)((n, \infty)) \leq \frac{\langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(\omega, t) \rangle}{n^{1+p}},$$

which is bounded by $\frac{M}{n^{1+p}}$ by the definition of $\Omega_B^r(M)$.

Note that $\bar{\mathcal{Z}}^{r,m}(\omega, t)(\mathbb{R}_+)$ is bounded by K^r/r by the policy constraint (2.8). By condition (4.12), $K^r/r \leq K + 1$ for all large r. The same argument applies for $\bar{\mathcal{Z}}^{r,m}(\omega, t)$. □

5.1.4 Asymptotic Regularity

Similar result as in this section was proved in Chapter 4. However, here we consider a much longer time horizon $[0, \lfloor rT \rfloor + L]$ instead of interval $[0, T]$. The proof of the following result use a combination of ideas in Chapter 4 and [24].

Lemma 5.5. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $T > 0$ *and* $L > 1$. *For each* $\epsilon, \eta > 0$ *there exists a* $\kappa > 0$ *(depending on* ϵ *and* η*) such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^{r,m}(t)([x, x+\kappa]) \le \epsilon \right) \ge 1-\eta. \tag{5.34}$$

Proof. To prove (5.34), it suffices to show

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,\lfloor rT\rfloor+L]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^{r,0}(t)([x, x+\kappa]) \le \epsilon \right) \ge 1-\eta.$$

First, We have that for any $\epsilon, \eta > 0$, there exists a κ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^{r,0}(0)([x, x+\kappa]) \le \epsilon/2 \right) \ge 1-\eta/2. \tag{5.35}$$

The proof of this inequality is exactly the same as the proof of (5.14) in Chapter 4, so we omit it for brevity.

Now we need to extend this result to the interval $[0, \lfloor rT \rfloor + L]$. Denote the event in (5.35) by Ω_1^r. Let

$$\Omega_2^r(M) = \Omega_1^r \cap \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M). \tag{5.36}$$

By (5.25), (5.27) and (5.33), there exists an $M > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_2^r(M)\right) \ge 1-\eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega_2^r(M)$.

For any $r > 0$, $t \in [0, T]$ we define the random time

$$t_0 = \sup\Big\{ \{ s \le t : \langle 1, \bar{\mathcal{Z}}^{r,0}(s) \rangle < \epsilon/4 \} \cup \{0\} \Big\}.$$

Let

$$t_1 = \max(t_0, t - \frac{2MK}{\epsilon}).$$

We have the following three cases for discussion.

If $t_1 = 0$, then by (5.35) for each $x \in \mathbb{R}_+$

$$\bar{\mathcal{Z}}^{r,0}(t_1)([x, x+\kappa] + \bar{S}^{r,0}(t_1, t)) \leq \epsilon/2.$$

If $t_1 = t_0$, then for each $\delta > 0$ there exists an s such that $t_1 - \delta < s < t_1$ and $\bar{Z}^{r,0}(s)(\mathbb{R}_+) < \epsilon/4$. Since we are only concerned with small ϵ (which should be small enough such that $\bar{Z}^{r,0}(s) < \epsilon/4 < K^r/r$), $\bar{Q}^{r,0}(s) = 0$ by the policy constraint (2.8). Note that (2.3) implies

$$\bar{B}^{r,0}(s,t) \leq \bar{E}^{r,0}(s,t) + \bar{Q}^{r,0}(s) \text{ for all } s \leq t. \tag{5.37}$$

By (5.24), we have $\bar{B}^{r,0}(s, t_1) \leq \lambda\delta + \epsilon_1$. For any Borel set A, by the fluid scaled system dynamic equation (5.21),

$$\bar{\mathcal{Z}}^{r,0}(t_1)(A) \leq \bar{\mathcal{Z}}^{r,0}(s)(\mathbb{R}_+) + \bar{B}^{r,0}(s, t_1) \leq \epsilon/4 + \lambda\delta + \epsilon_1,$$

which can be made smalled than $\epsilon/2$ by choosing ϵ_1, δ suitably small.

It $t_1 = t - \frac{2MK}{\epsilon}$, then $\bar{S}^{r,m}(t_1, t) \geq \frac{2M}{\epsilon}$. So

$$\bar{\mathcal{Z}}^{r,0}(t_1)([x, x+\kappa] + \bar{S}^{r,0}(t_1, t)) \leq \bar{\mathcal{Z}}^{r,0}(t_1)([\frac{2M}{\epsilon}, \infty)) \leq \epsilon/2$$

where the last inequality is due to the Markov's inequality and the definition of $\Omega^r_B(M)$. To summarize, we have

$$\bar{\mathcal{Z}}^{r,0}(t_1)([x, x+\kappa] + \bar{S}^{r,0}(t_1, t)) \leq \epsilon/2. \tag{5.38}$$

By the fluid scaled stochastic dynamic equation (5.21),

$$\begin{aligned}\bar{\mathcal{Z}}^{r,0}(t)([x, x+\kappa]) = \bar{\mathcal{Z}}^{r,0}(t_0)([x, x+\kappa] + \bar{S}^{r,0}(t_0, t)) \\ + \frac{1}{r} \sum_{i=r\bar{B}^{r,0}(t_0)+1}^{r\bar{B}^{r,0}(t)} \delta_{B^r_i}([x, x+\kappa] + \bar{S}^{r,0}(\tau_i, t)),\end{aligned}$$

for each $x \in \mathbb{R}_+$. By the choice of t_1, the first term on the right hand side of the above equation is always upper bounded by $\epsilon/2$. Let I denote the second term on the right hand side of the proceeding equation. Now it only remains to show that $I < \epsilon/2$.

Let $t_1, t_2, \cdots, t_N = t$ be a partition of the interval $[t_1, t]$ such that $|t_{j+1} - t_j| < \delta$ for all $j = 1, \cdots, N-1$, where δ and N are to be chosen below. Write I as the summation

$$I = \sum_{j=1}^{N-1} \frac{1}{r} \sum_{i=r\bar{B}^{r,0}(t_j)+1}^{r\bar{B}^{r,0}(t_{j+1})} \delta_{B_i^r}([x, x+\kappa] + \bar{S}^{r,0}(\tau_i, t)).$$

Note that by the definition of t_1,

$$N \leq \frac{2MK}{\delta\epsilon}. \tag{5.39}$$

Recall that τ_i^r is the time that the ith job starts service, so on each sub-interval $[t_j, t_{j+1}]$ those i's to be summed must satisfy $t_j \leq \tau_i^r \leq t_{j+1}$. This implies that

$$\bar{S}^{r,0}(t_{j+1}, t) \leq \bar{S}^{r,0}(\tau_i, t) \leq \bar{S}^{r,0}(t_j, t).$$

By the definition of t_1, we have $\bar{Z}^{r,0}(s) \geq \epsilon/4$ for all $s \in [t_1, t]$. So

$$\bar{S}^{r,0}(t_j, t_{j+1}) \leq \frac{4\delta}{\epsilon}.$$

Let

$$C_j = [x + \bar{S}^{r,0}(t_{j+1}, t), \quad x + \bar{S}^{r,0}(t_{j+1}, t) + \kappa + \frac{2\delta}{\epsilon}].$$

Then

$$I \leq \sum_{j=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{B}^{r,0}(t_j)+1}^{r\bar{B}^{r,0}(t_{j+1})} \delta_{B_i^r}(C_j).$$

By (5.24), (5.37) and the definition of $\Omega_B^r(M)$, we have for all $j = 0, \cdots, N-1$

$$-rM \leq r\bar{B}^{r,0}(t_j) \leq r(\lambda T + \epsilon_1 + M) \leq 2\lambda r T + rM,$$

$$\bar{B}^{r,0}(t_j, t_{j+1}) \leq \lambda T + \epsilon_1 + M \leq 2\lambda T + M.$$

Thus by (5.27),

$$\Big|\frac{1}{r} \sum_{i=r\bar{B}^{r,0}(t_j)+1}^{r\bar{B}^{r,0}(t_{j+1})} \delta_{B_i^r}(C_j) - \Big(\bar{B}^{r,0}(t_{j+1}) - \bar{B}^{r,0}(t_j)\Big)\nu^r(C_j)\Big| < \epsilon_1.$$

So

$$I \leq \sum_{j=0}^{N-1} [\bar{B}^{r,0}(t_{j+1}) - \bar{B}^{r,0}(t_j)]\nu^r(C_j) + N\epsilon_1.$$

By (4.9), for all $\epsilon_2 > 0$

$$\nu^r(C_j) \leq \nu(C_j) + \epsilon_2$$

for all large enough r. Since C_j is a close interval with length $\kappa + \frac{2\delta}{\epsilon}$, by (the condition that ν has not atoms) we can choose κ, δ small enough such that

$$\nu(C_j) < \frac{\epsilon}{4(2\lambda T + M)}.$$

Thus, we conclude that

$$\begin{aligned} I &\leq (\epsilon_2 + \frac{\epsilon}{4(2\lambda T + M)}) \sum_{j=0}^{N-1} [\bar{B}^{r,0}(t_{j+1}) - \bar{B}^{r,0}(t_j)] + N\epsilon_1 \\ &\leq (\epsilon_2 + \frac{\epsilon}{4(2\lambda T + M)})[\bar{B}^{r,0}(t) - \bar{B}^{r,0}(t_0)] + N\epsilon_1 \\ &\leq \epsilon_2(2\lambda T + M) + \frac{\epsilon}{4} + \frac{2MK}{\delta\epsilon\epsilon_1}, \end{aligned}$$

where the last inequality is again due to (5.35), (5.37) and (5.39). Finally, by choosing ϵ_1, ϵ_2 small enough, we obtain that $I < \epsilon/2$. □

5.1.5 Oscillation Bound

Consider a *càdlàg* function $\zeta(\cdot)$ on a fixed interval $[0, L]$ taking values in a metric space $(\mathbf{E}, \pi)$. The *modulus of continuity* is defined to be

$$\mathbf{w}_L(\zeta(\cdot), \delta) = \sup_{s,t \in [0,L], |s-t| < \delta} \pi[\zeta(s), \zeta(t)].$$

If the metric space is $\mathbb{R}$, we just use the Euclidian norm; if the space is $\mathbf{M}$ or $\mathbf{M}_1 \times \mathbf{M}_2$, we use the Prohorov metric $\mathbf{d}$ defined in Chapter 1. We have the following bound on the oscillation of the shifted fluid scaled measure valued processes.

Lemma 5.6. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $T > 0$ *and* $L > 1$. *For each* $\epsilon, \eta > 0$ *there exists a* $\delta > 0$ *(depending on* ϵ *and* η*) such that*

$$\liminf_{r \to \infty} \mathbb{P}\left(\max_{m \leq \lfloor rT \rfloor} \max\left(\mathbf{w}_L(\bar{\mathcal{Q}}^{r,m}(\cdot), \delta), \mathbf{w}_L(\bar{\mathcal{Z}}^{r,m}(\cdot), \delta) \right) \leq \epsilon \right) \geq 1 - \eta. \tag{5.40}$$

The proof of this lemma, which builds on the asymptotic regularity (Lemma 5.5), using the exactly same argument to prove Lemma 5.6 from Lemma 5.5 in Chapter 4. We omit this proof for brevity.

For any sequences $\{\kappa_i\}$ and $\{\delta_i\}$ consider the following set

$$\begin{aligned}&\Big\{ \max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^r(t)([x,x+\kappa_j]) \le \frac{1}{j}\Big\}\\ \cap \big\{ &\max_{m\le\lfloor rT\rfloor} \max\left(\mathbf{w}_L(\bar{\mathcal{Q}}^r(\cdot),\delta_j), \mathbf{w}_L(\bar{\mathcal{Z}}^r(\cdot),\delta_j)\right) \le \frac{1}{j}\Big\}.\end{aligned} \tag{5.41}$$

Denote the two sequences $\{\kappa_i\}$ and $\{\delta_i\}$ by $\mathcal{S}$. To emphasize the dependency on $\mathcal{S}$ and j, denote the above event by $\Omega_R^r(\mathcal{S},j)$. By Lemma 5.5 and Lemma 5.6, for any $\eta>0$, there exists an $\mathcal{S}$ such that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_R^r(\mathcal{S},j)\right) \ge 1-\frac{\eta}{2^j} \quad \text{for } j=1,2,\cdots. \tag{5.42}$$

Now define

$$\Omega_R^r(\mathcal{S}) = \bigcap_{j=1}^{\lfloor r\rfloor} \Omega_R^r(\mathcal{S},j). \tag{5.43}$$

It follows from (5.42) that

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega_R^r(\mathcal{S})\right) \ge 1-\eta. \tag{5.44}$$

Denote

$$\Omega^r(M,\mathcal{S}) = \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M+2\lambda T, M+2\lambda L) \cap \Omega_R^r(\mathcal{S}). \tag{5.45}$$

We are now ready to present the precompactness result.

Theorem 5.3. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $T>0$ *and* $L>1$. *For each* $\eta>0$, *the exists a constant* $M>0$ *and an* $\mathcal{S}$ *such that such that*

$$\liminf_{r\to\infty} \mathbb{P}\left(\Omega^r(M,\mathcal{S})\right) \ge 1-\eta. \tag{5.46}$$

Suppose $\{r_n\}_{n\in\mathbb{N}}$ *is a sequence in* $\mathbb{R}_+$ *which goes to infinity. Any sequence of functions* $\{(\bar{\mathcal{Q}}^{r_n,m_n}(\omega_n,\cdot), \bar{\mathcal{Z}}^{r_n,m_n}(\omega_n,\cdot))\}_{n\in\mathbb{N}}$ *with* $\omega_n \in \Omega^{r_n}(M,\mathcal{S})$ *and* $m_n \le \lfloor r_n T\rfloor$ *for each*

$n \in \mathbb{N}$ has a subsequence $\{(\bar{\mathcal{Q}}^{r_{n_i},m_{n_i}}(\omega_{n_i},\cdot),\bar{\mathcal{Z}}^{r_{n_i},m_{n_i}}(\omega_{n_i},\cdot))\}_{i\in\mathbb{N}}$ such that

$$\upsilon_T[(\bar{\mathcal{Q}}^{r_{n_i},m_{n_i}}(\omega_{n_i},t),\bar{\mathcal{Z}}^{r_{n_i},m_{n_i}}(\omega_{n_i},t)),(\tilde{\mathcal{Q}}(t),\tilde{\mathcal{Z}}(t))] \to 0 \quad \textit{as } i \to \infty,$$

for some process $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))$ which is continuous, where υ_T is the uniform metric defined in (1.2).

Proof. The probability inequality follows immediately from (5.25), (5.33), (5.27) and (5.44).

The space $\mathbf{M}_1\times\mathbf{M}_2$ endowed with the metric $\mathbf{d}$ (defined in Section 1.4) is complete. Lemma 5.4 verifies condition (a) in Theorem 3.6.3 of [17]. For any $\epsilon > 0$ there exists a j_0 such that $1/j < \epsilon$ for all $j > j_0$. By (5.41) and (5.43), we have that when $\delta \le \delta_{j_0}$ and $r > j_0$,

$$\max\left(\mathbf{w}_T(\bar{\mathcal{Q}}^{r,m}(\omega^r,\cdot),\delta),\mathbf{w}_T(\bar{\mathcal{Z}}^{r,m}(\omega^r,\cdot),\delta)\right) < \epsilon, \tag{5.47}$$

for any $\omega^r \in \Omega^r(M,\mathcal{S})$ and $m \le \lfloor rT \rfloor$. This verifies condition (b) in Theorem 3.6.3 of [17]. So the sequence $\{(\bar{\mathcal{Q}}^{r_n,m_n}(\omega^{r_n},\cdot),\bar{\mathcal{Z}}^{r_n,m_n}(\omega^{r_n},\cdot))\}_{n\in\mathbb{N}}$ is precompact in the space $\mathbf{D}([0,T],\mathbf{M}_1\times\mathbf{M}_2)$ endowed with the Skorohod J_1 topology. In other words, there is a convergent subsequence. The limit of this subsequence is continuous by the oscillation bound (5.47). So convergence in the Skorohod J_1-topology is the same as convergence in the uniform metric defined in Section 1.4. □

5.2 State Space Collapse

In this section, we introduce the set of fluid limits. We prove each fluid limit is a fluid model solution. We then show that the set of fluid limits is "rich". Finally, we present the proof of the state space collapse result, Theorem 5.2.

5.2.1 Fluid Limits

Let $\mathscr{D}_L(M,\mathcal{S})$ denote the set of fluid limits of all convergent subsequences of sequences as in Theorem 5.3. It is then quite clear that we have the following property.

Lemma 5.7. *Assume* (4.9)–(4.15), (5.5)–(5.9). *The set of fluid limits* $\mathscr{D}_L(M,\mathcal{S})$ *is non-empty. Pick an element* $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M,\mathcal{S})$, *for any* $\epsilon > 0$ *and* $r_0 \in \mathbb{R}_+$, *there exists an* $r \geq r_0$, $m \leq \lfloor rT \rfloor$ *and* $\omega \in \Omega^r(M,\mathcal{S})$ *such that*

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,m}(\omega,\cdot),\bar{\mathcal{Z}}^{r,m}(\omega,\cdot)),(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))] < \epsilon.$$

Roughly speaking, this lemma says that any element in $\mathscr{D}_L(M,\mathcal{S})$ can be approximated by a shifted fluid scaled process of the rth system evaluated at some sample path in $\Omega^r(M,\mathcal{S})$ with arbitrarily large index r. This helps prove the following property of the fluid limits.

Fix a constant $0 < q < p$, where p is the same one as in (4.10) and (4.14). Recall the subset $\mathscr{I}^q_{2M}$ of all valid initial conditions defined in (3.42).

Lemma 5.8. *Assume* (4.9)–(4.15), (5.5)–(5.9). *Fix* $L > 0$ *and* $0 < q < p$. *Any element* $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M,\mathcal{S})$ *is a critically loaded fluid model solution with initial condition belongs to* $\mathscr{I}^q_{2M}$.

Proof. We first prove the initial condition $(\tilde{\mathcal{Q}}(0),\tilde{\mathcal{Z}}(0)) \in \mathscr{I}^q_{2M}$. By the definition of the fluid limit, there exists a subsequence

$$(\bar{\mathcal{Q}}^{r_i,m_i}(\omega_i,0),\bar{\mathcal{Z}}^{r_i,m_i}(\omega_i,0)) \to (\tilde{\mathcal{Q}}(0),\tilde{\mathcal{Z}}(0)) \quad \text{as } i \to \infty,$$

where the above convergence is in the Prohorov metric. By (5.33) and (5.45), we have

$$\langle 1, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i,0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i,0)\rangle < M,$$
$$\langle \chi^p, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i,0), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i,0)\rangle < M.$$

This implies that,

$$\begin{aligned}
&\langle \chi^p, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i,0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i,0)\rangle \\
&\leq \langle 1, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i,0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i,0)\rangle + \langle \chi^p, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i,0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i,0)\rangle \\
&\leq 2M.
\end{aligned}$$

By the corollary of Theorem 25.12 in [4], we have that for any $0 < q < p$

$$\langle \chi^q, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle \to \langle \chi^q, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle,$$
$$\langle \chi, \bar{\mathcal{Q}}^{r_i}(0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle \to \langle \chi, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle,$$

as $i \to \infty$. This implies that $\langle \chi^q, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle < 2M$ and $\langle \chi, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle < M$, which yields the result.

By Lemma 5.7, for any fluid limit $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ can be approximated by a shifted fluid scaled process of the rth process evaluated at some sample path in $\Omega^r(M, \mathcal{S})$ with arbitrarily large index $r \in \mathbb{R}_+$, it satisfies the stochastic dynamic equations (2.5) and (2.6). It then follows from the same argument Lemmas 4.8–4.10 in Chapter 4 that each fluid limit satisfies the fluid model equations (3.9) and (3.10) and constraints (3.11)–(3.13). □

5.2.2 Uniform Approximation

The following lemma is analogous to Lemma 4.1 in [7].

Lemma 5.9. *Assume* (4.9)–(4.15), (5.5)–(5.9). *For each* $\epsilon > 0$, *there exists an* $r_0 \in \mathbb{R}_+$ *such that for any* $r \geq r_0$, $m \leq \lfloor rT \rfloor$ *and* $\omega \in \Omega^r(M, \mathcal{S})$, *we can find a* $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ *satisfying*

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,m}(\omega, \cdot), \bar{\mathcal{Z}}^{r,m}(\omega, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] < \epsilon.$$

Proof. Assume it is not true. Then there exists an $\epsilon > 0$ such that for any natural number i there exist an $r_i > i$, $m_i \in \lfloor rT \rfloor$ and $\omega_i \in \Omega^r(M, \mathcal{S})$ such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, \cdot), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] \geq \epsilon,$$

for all $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$. However, by Theorem 5.3, the sequence

$$\{(\bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, \cdot), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, \cdot))\}_{i=0}^{\infty}$$

contains a convergent subsequence, the limit of which must be in $\mathscr{D}_L(M, \mathcal{S})$. This is a contradiction. □

In contrast to Lemma 5.7, this lemma says that any shifted fluid scaled process of the rth system evaluated at some sample path in $\Omega^r(M,\mathcal{S})$ with index r large enough can be approximated by some element in $\mathscr{D}_L(M,\mathcal{S})$, which has been proved to be fluid model solution in Lemma 5.8. This result will help prove the state space collapse result for diffusion scaled processes.

5.2.3 Proof of State Space Collapse

Proof of Theorem 5.2. By (5.46), it suffices to show that for each $\epsilon > 0$, there exists an r_0 such that when $r > r_0$,

$$\sup_{\omega\in\Omega^r(M,\mathcal{S})}\sup_{t\in[0,T]}\mathbf{d}[(\hat{\mathcal{Q}}^r(\omega,t),\hat{\mathcal{Z}}^r(\omega,t)),\Delta_{K,\lambda}\hat{W}^r(\omega,t)] < \epsilon. \tag{5.48}$$

By Lemma 5.8, any $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M,\mathcal{S})$ is a critically loaded fluid model solution with initial condition $(\xi,\mu) \in \mathscr{I}^q_{2M}$. Denote

$$\tilde{W}(\cdot) = \langle \chi, \tilde{\mathcal{Q}}(\cdot) + \tilde{\mathcal{Z}}(\cdot)\rangle.$$

It follows from the workload conservation property in Proposition 3.1 that $\tilde{W}(\cdot) \equiv \langle \chi, \xi+\mu\rangle$. By Theorem 3.4, there exists an $L^* > 0$ such that when $s > L^*$,

$$\mathbf{d}[(\tilde{\mathcal{Q}}(s),\tilde{\mathcal{Z}}(s)),\Delta_{K,\nu}\tilde{W}(s)] < \epsilon/3, \tag{5.49}$$

for all $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M,\mathcal{S})$. Now, fix a constant $L > L^* + 1$. Note that

$$[0, r^2T] \subset [0,L^*] \bigcup_{m=0}^{\lfloor rT\rfloor} [mr + L^*, mr + L].$$

By the definition of diffusion and shifted fluid scaling, to show (5.48) it suffices to show

$$\sup_{\omega\in\Omega^r(M,\mathcal{S})}\max_{m\le\lfloor rT\rfloor}\sup_{s\in[L^*,L]}\mathbf{d}[(\bar{\mathcal{Q}}^{r,m}(\omega,s),\bar{\mathcal{Z}}^{r,m}(\omega,s)),\Delta_{K,\lambda}\bar{W}^{r,m}(\omega,s)] < \epsilon, \tag{5.50}$$

$$\sup_{\omega\in\Omega^r(M,\mathcal{S})}\sup_{s\in[0,L^*]}\mathbf{d}[(\bar{\mathcal{Q}}^{r,0}(\omega,s),\bar{\mathcal{Z}}^{r,0}(\omega,s)),\Delta_{K,\lambda}\bar{W}^{r,0}(\omega,s)] < \epsilon. \tag{5.51}$$

We first prove (5.50). By Lemma 5.9, for any $\epsilon' > 0$, there exists a $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ (depending on r, m and ω) such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,m}(\omega,\cdot), \bar{\mathcal{Z}}^{r,m}(\omega,\cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] < \epsilon'.$$

for any $r > r_0$, $\omega \in \Omega^r(M, \mathcal{S})$ and $m \leq \lfloor rT \rfloor$. By the definition of $\Omega^r(M, \mathcal{S})$ and Proposition 5.2, we have that for each fixed $0 < q < p$, both $\langle \chi^{1+q}, \tilde{\mathcal{Q}}(\cdot) + \tilde{\mathcal{Z}}(\cdot) \rangle$ and $\langle \chi^{1+q}, \bar{\mathcal{Q}}^{r,m}(\omega,\cdot) + \bar{\mathcal{Z}}^{r,m}(\omega,\cdot) \rangle$ are uniformly bounded. It then follows from Lemma C.2 and by taking ϵ' small enough that

$$\sup_{t\in[0,L]} |\tilde{W}(t) - \bar{W}^{r,m}(\omega,t)| < \epsilon/3. \tag{5.52}$$

Note that for any real numbers w_1, w_2, by the definition of the lifting map $\Delta_{K,\nu}$, we have

$$\mathbf{d}[\Delta_{K,\nu} w_1, \Delta_{K,\nu} w_2] < |w_1 - w_2|. \tag{5.53}$$

So (5.50) follows from (5.49) and the above three inequalities.

It now remains to show (5.51). By Lemma 5.9, for any $\epsilon' > 0$, there exists a $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ (depending on r and ω) such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,0}(\omega,\cdot), \bar{\mathcal{Z}}^{r,0}(\omega,\cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] < \epsilon'. \tag{5.54}$$

By conditions (4.13) and (5.9), we have that

$$(\tilde{\mathcal{Q}}(0), \tilde{\mathcal{Z}}(0)) = \Delta_{K,\nu} \tilde{W}(0).$$

In other words, the initial condition $(\tilde{\mathcal{Q}}(0), \tilde{\mathcal{Z}}(0))$ is an equilibrium state. Since $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is a fluid model solution, by Theorem 3.4,

$$(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t)) = \Delta_{K,\nu} \tilde{W}(t) \quad \text{for all } t \geq 0.$$

So (5.51) follows immediately from (5.53), (5.54) and the above equation. □

CHAPTER VI

STEADY STATE OF LIMITED PROCESSOR SHARING QUEUES

It is clear that the measure-valued process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$ for the LPS queue is a regenerative process. Let $R_0 = 0$ and define the regenerative points R_n, $n \geq 1$ as the following:

$$R_n = \inf\big\{t > R_{n-1} : W(t-) = 0 \text{ and } W(t) > 0\big\}. \tag{6.1}$$

The regeneration points are those time epochs that the workload jumps from 0. It is clear that the jump happens because of the new arrival. By the definition of workload, $(\mathcal{Q}(t), \mathcal{Z}(t)) = (\mathbf{0}, \mathbf{0})$ if and only if $W(t) = 0$ for any $t \geq 0$. So the process starts from empty at time R_n with a new job just arrives at R_n. Thus, the evolution of the process from time R_n onwards does not depend on any information of the process before that time.

Note that the workload process of a single buffer single server system is the same for all non-idling policies. It is well-known that the workload process (for any non-idling policy) is a delayed regenerative process if $W(0) > 0$, and the above definition of R_n is one way to define the regenerative points. By Proposition 3.1 in Chapter X of [2], the mean of the regenerative cycles Y_i's ($Y_i = R_i - R_{i-1}$) with $i > 1$ is finite if $\rho < 1$. By Proposition 3.2 in Chapter X of [2], the distribution of them is non-lattice if the service time distribution F is non-lattice.

In summary, the process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$ can be modeled as a delayed regenerative process. Denote $\mathbb{E}_0(\cdot) = \mathbb{E}(\cdot|(\mathcal{Q}(0), \mathcal{Z}(0)) = (\mathbf{0}, \delta_{v_1}), U_1 = 0)$, that is, the expectation operator given that the queueing process start from empty and there is an arrival at time 0. We write $Y = Y_1$ for the length of the first cycle. Now, we define a

distribution π on the polish space $\mathbf{M}_1 \times \mathbf{M}_2$ by

$$\pi(A) = \frac{1}{\mathbb{E}_0 Y} \mathbb{E}_0 \int_0^Y 1_{\{(\mathcal{Q}(t), \mathcal{Z}(t)) \in A\}} ds,$$

for any Borel set $A \in \mathbf{M}_1 \times \mathbf{M}_2$. The following result about the steady state distribution of the LPS queue follows directly from Theorem 1.2 in Chapter X of [2].

Proposition 6.1 (Stochastic Stability of LPS)**.** *Suppose that the traffic intensity $\rho < 1$ and the service time distribution F is non-lattice. The above defined distribution π is the unique stationary distribution for the measure-valued process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$. The distribution of $(\mathcal{Q}(t), \mathcal{Z}(t))$ converges to π as $t \to \infty$.*

The above theorem establishes the convergence of the regenerative process to the steady state limit, which has the stationary distribution π. In fact, there exists a stationary version $(\mathcal{Q}_\pi(\cdot), \mathcal{Z}_\pi(\cdot))$ of the regenerative process (see [51]) such that the marginal distribution at any time $t \geq 0$ is π. The stationarity of the process $(\mathcal{Q}_\pi(\cdot), \mathcal{Z}_\pi(\cdot))$ will help to obtain a coupling inequality later.

6.1 Validity of Heavy Traffic Steady State Approximations

As we see in Theorem 5.1, the heavy traffic limiting process $(\mathcal{Q}^*(\cdot), \mathcal{Z}^*(\cdot))$ is the image of the workload process $W^*(\cdot)$ under the continuous mapping $\Delta_{K,\nu}$. The limit is in the sense of weak convergence of probability measures, so the limiting process may not be in the same probability space where each process with index r is defined. Denote $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ the probability space where the weak limit is defined. It is well-known that the marginal distribution $W^*(t)$ of the reflected Brownian motion $W^*(\cdot)$ converges weakly to that of the steady state random variable $W^*(\infty)$, which has the stationary distribution,

$$\mathbb{P}^* \left(W^*(\infty) > x \right) = \exp\left(\frac{-2\theta x}{\beta(c_a^2 + c_s^2)} \right). \tag{6.2}$$

By the continuous mapping theorem, the measure-valued process $(\mathcal{Q}^*(\cdot), \mathcal{Z}^*(\cdot))$ converges weakly to $\Delta_{K,\nu} W^*(\infty)$. Denote the distribution of $\Delta_{K,\nu} W^*(\infty)$ by π^*. For

any open set $B \in \mathbf{M}_1 \times \mathbf{M}_2$,

$$\pi^*(B) = \mathbb{P}^* \left(\Delta_{K,\nu} W^*(\infty) \in B \right) = \mathbb{P}^* \left(W^*(\infty) \in \Delta_{K,\nu}^{-1} B \right). \tag{6.3}$$

On the other hand, for each r, since the traffic intensity $\rho^r < 1$ and the service time distribution is non-lattice, the diffusion scaled process $(\hat{\mathcal{Q}}^r(\cdot), \hat{\mathcal{Z}}^r(\cdot))$ is a regenerative process. By Proposition 6.1, $(\hat{\mathcal{Q}}^r(t), \hat{\mathcal{Z}}^r(t))$ converges to the steady state $(\hat{\mathcal{Q}}^r(\infty), \hat{\mathcal{Z}}^r(\infty))$ which has distribution $\hat{\pi}^r$ as $t \to \infty$.

Now, the question is: does the stationary distribution $\hat{\pi}^r$ converges to π^*, which is obtained by first taking heavy traffic limit and then steady state limit? We have the following theorem, which validates the interchange of steady state limit and heavy traffic limit.

Theorem 6.1. *Assume* (4.9)–(4.15), (5.5)–(5.9). *The sequence* $\{\hat{\pi}^r\}$ *converges weakly to* π^*.

The major steps of proving the above theorem are, first, obtaining inequality (6.7) via coupling, and second, establishing the uniform convergence on the right hand side of (6.7) (i.e. the uniform bound of the coupling time) in Lemma 6.1. The proof of Theorem 6.1 will be presented at the end of this section.

Following the discussion in at the beginning of this chapter, we can construct a stationary version of the regenerative process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ such that at any time $t \geq 0$, it has distribution $\hat{\pi}^r$, i.e.

$$\mathbb{P}\left((\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(t), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(t)) \in B \right) = \hat{\pi}^r(B), \tag{6.4}$$

for any open set $B \in \mathbf{M}_1 \times \mathbf{M}_2$. Let $\hat{W}^r_{\hat{\pi}^r}(\cdot) = \langle \chi, \hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot) + \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot) \rangle$ denote the corresponding workload process.

Let us now couple the stationary process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ with the corresponding process $(\hat{\mathcal{Q}}^r_0(\cdot), \hat{\mathcal{Z}}^r_0(\cdot))$ which starts with a zero initial condition. In other words, both $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ and $(\hat{\mathcal{Q}}^r_0(\cdot), \hat{\mathcal{Z}}^r_0(\cdot))$ are driven by the same stochastic primitives

$(\hat{E}^r(\cdot), \{v_i^r\}_{i\geq 1})$. The only difference is the initial condition. Note that the stationarity assumption forces the renewal arrival process to be a stationary delayed renewal process. Define

$$\hat{t}_c^r = \inf\left\{t \geq 0 : (\hat{\mathcal{Q}}_{\hat{\pi}^r}^r(t), \hat{\mathcal{Z}}_{\hat{\pi}^r}^r(t)) = (\mathbf{0}, \mathbf{0})\right\}. \tag{6.5}$$

Note that the workload of $(\hat{\mathcal{Q}}_0^r(\cdot), \hat{\mathcal{Z}}_0^r(\cdot))$ starts at 0, which is less than or equal to $\hat{W}_{\hat{\pi}^r}^r(0)$. Since the LPS policy is work conserving, and both processes have the same stochastic primitives, for any $t \geq 0$

$$(\hat{\mathcal{Q}}_{\hat{\pi}^r}^r(t), \hat{\mathcal{Z}}_{\hat{\pi}^r}^r(t)) = (\mathbf{0}, \mathbf{0}) \text{ implies } (\hat{\mathcal{Q}}_0^r(t), \hat{\mathcal{Z}}_0^r(t)) = (\mathbf{0}, \mathbf{0}). \tag{6.6}$$

Since both systems are driven by the same arrival process, $(\hat{\mathcal{Q}}_{\hat{\pi}^r}^r(t), \hat{\mathcal{Z}}_{\hat{\pi}^r}^r(t))$ and $(\hat{\mathcal{Q}}_0^r(t), \hat{\mathcal{Z}}_0^r(t))$ are identical for all $t \geq t_c$. It then follows from Corollary 2.2 in Chapter VII of [2] that

$$\left|\mathbb{P}\left((\hat{\mathcal{Q}}_0^r(t), \hat{\mathcal{Z}}_0^r(t)) \in B\right) - \hat{\pi}^r(B)\right| \leq \mathbb{P}\left(\hat{t}_c^r > t\right) \quad \text{for all } t \geq 0. \tag{6.7}$$

We now show that the probability $\mathbb{P}\left(\hat{t}_c^r > t\right)$ converges to 0 as $t \to \infty$ uniformly in r.

Lemma 6.1. *Assume* (4.9)–(4.15), (5.5)–(5.9). *We have that*

$$\sup_r \mathbb{P}\left(\hat{t}_c^r > t\right) \to 0 \text{ as } t \to \infty. \tag{6.8}$$

Proof. Let $\hat{C}^r(t) = \frac{1}{r}\sum_{i=1}^{E^r(r^2 t)} v_i^r - rt$ for all $t \geq 0$. The summation in the above denotes the total amount of arrived work (under diffusion scaling) by time t, the second term $-rt$ denote the amount of work the server has finished by time t without idling. So the first time the process $(\hat{\mathcal{Q}}_{\hat{\pi}^r}^r(\cdot), \hat{\mathcal{Z}}_{\hat{\pi}^r}^r(\cdot))$ reaches zero is the first time that $\hat{W}_{\hat{\pi}^r}^r(0) + \hat{C}^r(t) = 0$. By the definition of $\hat{t}_c^r$ in (6.5),

$$\hat{t}_c^r = \inf\{t \geq 0 : \hat{C}^r(t) = -\hat{W}_{\hat{\pi}^r}^r(0)\}.$$

So for any $M > 0$,

$$\begin{aligned}
\mathbb{P}\left(\hat{t}_c^r > t\right) &= \mathbb{P}\left(\hat{C}^r(s) > -\hat{W}^r_{\hat{\pi}^r}(0), \text{ for all } s \le t\right) \\
&\le \mathbb{P}\left(\hat{C}^r(t) > -\hat{W}^r_{\hat{\pi}^r}(0)\right) \\
&\le \mathbb{P}\left(\hat{C}^r(t) > -\hat{W}^r_{\hat{\pi}^r}(0), -\hat{W}^r_{\hat{\pi}^r}(0) \ge -M\right) + \mathbb{P}\left(-\hat{W}^r_{\hat{\pi}^r}(0) < -M\right) \\
&\le \mathbb{P}\left(\hat{C}^r(t) > -M\right) + \mathbb{P}\left(\hat{W}^r_{\hat{\pi}^r}(0) > M\right).
\end{aligned}$$

Since the regenerative process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ is stationary, the corresponding workload process $\hat{W}^r_{\hat{\pi}^r}(\cdot)$ is also stationary. By Corollary 7.5 in Chapter X of [2], the stationary distribution of the workload converges weakly to an exponential distribution as $r \to \infty$. This implies that for any $\epsilon > 0$, there exists $M' > 0$ such that

$$\sup_r \mathbb{P}\left(\hat{W}^r_{\hat{\pi}^r}(0) > M\right) < \epsilon \quad \text{for all } M \ge M'. \tag{6.9}$$

From now on, we fix a constant $M \in [M', \infty)$. Note that the process $\hat{C}^r(\cdot)$ converges weakly to a Brownian motion $B(\cdot)$ with drift $-\theta$ and variance $\beta(c_a^2 + c_s^2)$. This implies that

$$\lim_r \mathbb{P}\left(\hat{C}^r(t) > -M\right) = \mathbb{P}\left(B(t) > -M\right),$$

which goes to zero as $t \to \infty$. So for any constant M, there exists $t(M)$ such that

$$\limsup_r \mathbb{P}\left(\hat{C}^r(t) > -M\right) < \epsilon/2, \quad \text{for all } t \ge t(M).$$

This means that there exists $r_0 > 0$ such that for all $r \ge r_0$

$$\mathbb{P}\left(\hat{C}^r(t) > -M\right) < \epsilon, \quad \text{for all } t \ge t(M).$$

For each $r < r_0$, we can choose $t_r(M)$ large enough (depending on r) such that

$$\mathbb{P}\left(\hat{C}^r(t) > -M\right) < \epsilon, \quad \text{for all } t \ge t_r(M).$$

Since there are only finitely many of those r's that are less than r_0, let $t_0(M) = \max_{r<r_0} t_r(M)$. We now have that

$$\sup_r \mathbb{P}\left(\hat{C}^r(t) > -M\right) < \epsilon, \quad \text{for all } t \ge \max(t(M), t_0(M)). \tag{6.10}$$

The lemma follows immediately from (6.9) and (6.10). □

Proof of Theorem 6.1. For any closed set $B \in \mathbf{M}_1 \times \mathbf{M}_2$, we have

$$\begin{aligned}\hat{\pi}^r(B) - \pi^*(B) \leq & \left| \hat{\pi}^r(B) - \mathbb{P}\left((\hat{\mathcal{Q}}_0^r(t), \hat{\mathcal{Z}}_0^r(t)) \in B \right) \right| \\ & + \mathbb{P}\left((\hat{\mathcal{Q}}_0^r(t), \hat{\mathcal{Z}}_0^r(t)) \in B \right) - \mathbb{P}^*\left((\mathcal{Q}^*(t), \mathcal{Z}^*(t)) \in B \right) \\ & + \mathbb{P}^*\left((\mathcal{Q}^*(t), \mathcal{Z}^*(t)) \in B \right) - \pi^*(B).\end{aligned} \tag{6.11}$$

By the coupling inequality, the first term on the right hand side of in (6.11) is bounded by

$$\sup_r \mathbb{P}\left(\hat{t}_c^r > t \right),$$

which vanishes as $t \to \infty$, as proved in Lemma 6.1. According to the definition of π^* and Portmanteau Theorem (c.f. Theorem 2.1 in [5]), the lim sup of the third term on the right hand side of (6.11) equals to 0 as $t \to \infty$. So for any $\epsilon > 0$, there exists a $t_1 > 0$ (may be very large, but still finite) such that

$$\begin{aligned}\sup_r \mathbb{P}\left(\hat{t}_c^r > t_1 \right) < \epsilon, \\ \mathbb{P}^*\left((\mathcal{Q}^*(t_1), \mathcal{Z}^*(t_1)) \in B \right) - \pi^*(B) < \epsilon.\end{aligned}$$

For this fixed t_1, by Theorem 5.1 and Portmanteau Theorem, we have that

$$\limsup_r \mathbb{P}\left((\hat{\mathcal{Q}}_0^r(t_1), \hat{\mathcal{Z}}_0^r(t_1)) \in B \right) \leq \mathbb{P}^*\left((\mathcal{Q}^*(t_1), \mathcal{Z}^*(t_1)) \in B \right).$$

So there exists r_0 such that when $r \geq r_0$,

$$\mathbb{P}\left((\hat{\mathcal{Q}}_0^r(t_1), \hat{\mathcal{Z}}_0^r(t_1)) \in B \right) - \mathbb{P}^*\left((\mathcal{Q}^*(t_1), \mathcal{Z}^*(t_1)) \in B \right) < \epsilon.$$

So we have that for any $\epsilon > 0$, there exists r_0 such that when $r \geq r_0$,

$$\hat{\pi}^r(B) - \pi^*(B) < 3\epsilon.$$

This implies that $\limsup_r \hat{\pi}^r(B) \leq \pi^*(B)$ for any closed set B. The result of the theorem follows from Portmanteau Theorem. □

6.2 Performance Evaluation

So far, we have obtained results for the measure-valued description of the LPS queue. We now establish some more concrete results on the queue size, delay probability and response time.

6.2.1 Queue Length and Delay Probability

It follows from Corollary 5.1 that $X^*(\cdot)$ is a reflected Brownian motion with drift $\frac{-\theta}{\beta}$ and variance $\frac{c_a^2+c_s^2}{\beta}$ when it is above K and with drift $\frac{-\theta}{\beta_e}$ and variance $\frac{\beta(c_a^2+c_s^2)}{\beta_e^2}$ when it is below K. Define the map $f : \mathbb{R}_+ \to \mathbb{R}_+$ by $f(w) = \frac{1}{\beta}(w - K\beta_e)^+ + \frac{1}{\beta_e}(w \wedge K\beta_e)$ for all $w \in \mathbb{R}_+$. It is clear that $f(\cdot)$ is a continuous mapping with inverse

$$f^{-1}(w) = \begin{cases} \beta_e x & x \le K, \\ \beta_e K + \beta(x - K) & x > K. \end{cases}$$

By the continuous mapping theorem, $X^*(t)$ converges weakly to the steady state $X^*(\infty) = f(W^*(\infty))$ as $t \to \infty$, and $\mathbb{P}(X^*(\infty) > x) = \mathbb{P}(W^*(\infty) > f^{-1}(x))$. By the definition of equilibrium distribution, it is easy to see that $\frac{\beta_e}{\beta} = \frac{1+c_s^2}{2}$. Since the stationary distribution of the reflected Brownian motion $W^*(\cdot)$ is explicitly known as in (6.2), it is easy to compute the distribution of $X^*(\infty)$,

$$\mathbb{P}(X^*(\infty) > x) = \begin{cases} \exp\Big(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}x\Big) & x \le K, \\ \exp\Big(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}K - \frac{2\theta}{c_a^2+c_s^2}(x-K)\Big) & x > K. \end{cases} \tag{6.12}$$

Let

$$d_p^*(\infty) = \mathbb{P}(X^*(\infty) > K) = \exp\Big(-\frac{1+c_s^2}{c_a^2+c_s^2}\theta K\Big) \tag{6.13}$$

be the steady state probability that the limiting queue size $X^*(\cdot)$ is above the sharing level K. As we see, the steady state limit of the heavy traffic limit is so tractable that the stationary distribution can be explicitly written down. Since we have established the interchange of steady state limit and heavy traffic limit, the following result is a direct implication of Theorem 6.1 and the continuous mapping theorem.

Corollary 6.1. *Assume* (4.9)–(4.15), (5.5)–(5.9). *We have that*

$$\hat{X}^r(\infty) \Rightarrow X^*(\infty),$$
$$\hat{d}_p^r(\infty) \to d_p^*(\infty),$$

as $r \to \infty$, *where* $\hat{d}_p^r(\infty) = \mathbb{P}\left(\hat{X}^r(\infty) > K^r/r\right)$ *is the steady state delay probability for the* r*th system.*

6.2.2 Response Time

Let $R(t, v)$ denote the total time (including both waiting and service times) a job will stay in the system if it arrives at time t and with job size v. Since at a time t, there may not be an arrival or the arrival may not have job size v, the quantity $R(t, v)$ is often referred as the *virtual* response time. It contains two parts,

$$R(t, v) = R_B(t) + R_Z(t, v), \tag{6.14}$$

where $R_B(t)$ is the time that this virtual job spend waiting in buffer (which does not depend on its job size) and $R_S(t, v)$ is the service time of this virtual job. Let $W_B(\cdot) = \langle \chi, \mathcal{Q}(\cdot) \rangle$ and $W_Z(\cdot) = \langle \chi, \mathcal{Z}(\cdot) \rangle$ denote the workload in buffer and the workload in server respectively. From the time this virtual job enters the system t until it is about to enter service $t + R_B(t)$, the server never idles. So the workload the server processes during this time period is equal to $R_B(t)$. Since the LPS policy is workload conserving, we must have that

$$W_B(t) + W_Z(t) = R_B(t) + W_Z\big(t + R_B(t)\big). \tag{6.15}$$

It is clear that the service time of this virtual job should satisfy

$$S\Big(t + R_B(t), t + R_B(t) + R_Z(t, v)\Big) = v \tag{6.16}$$

We now study the heavy traffic limit of the diffusion scaled virtual response time $\hat{R}^r(t, v) = \frac{1}{r} R^r(r^2 t, v)$.

Proposition 6.2 (Heavy Traffic limit for virtual Response Time process)**.** *Assume* (4.9)–(4.15), (5.5)–(5.9). *For any fixed* $v \geq 0$*, the diffusion scaled virtual waiting time* $\left(\hat{R}_B^r(\cdot), \hat{R}_Z^r(\cdot, v)\right)$ *converges weakly to* $\left(R_B^*(\cdot), R_Z^*(\cdot, v)\right)$*, where*

$$R_B^*(t) = \beta(X^*(t) - K)^+, \quad R_Z^*(t, v) = v(X^*(t) \wedge K), \quad t \geq 0. \tag{6.17}$$

Proof. Since $\hat{W}_B^r(\cdot) = \langle \chi, \hat{\mathcal{Q}}^r(\cdot) \rangle$ and $\hat{W}_Z^r(\cdot) = \langle \chi, \hat{\mathcal{Z}}^r(\cdot) \rangle$, by Theorem 5.1 and the continuous mapping theorem,

$$(\hat{W}_B^r(\cdot), \hat{W}_Z^r(\cdot)) \Rightarrow \left((W^*(\cdot) - K\beta_e)^+, (W^*(\cdot) \wedge K\beta_e)\right), \tag{6.18}$$

as $r \to \infty$. The diffusion scaled version of the equation (6.15) can be written as

$$\hat{W}_B^r(t) + \hat{W}_Z^r(t) = \hat{R}_B^r(t) + \hat{W}_Z^r\big(t + \frac{1}{r}\hat{R}_B^r(t)\big). \tag{6.19}$$

It is clear that $\hat{R}_B^r(t) \leq \hat{W}^r(t)$, which converges to a reflected Brownian motion as $r \to \infty$. So on any finite interval $[0, T]$, for any $\epsilon > 0$ there exists $M > 0$ such that

$$\limsup_{r \to \infty} \mathbb{P}\left(\sup_{t \in [0,T]} \hat{R}_B^r(t) > M\right) < \epsilon.$$

Since $\hat{W}_Z^r(\cdot)$ converges to $W^*(\cdot) \wedge K\beta_e$, which is almost surely continuous, we have that

$$\sup_{t \in [0,T]} \left|\hat{W}_Z^r(t + \frac{1}{r}M) - \hat{W}_Z^r(t)\right| \Rightarrow 0 \text{ as } r \to \infty.$$

So for any $\epsilon > 0$,

$$\limsup_{r \to \infty} \mathbb{P}\left(\sup_{t \in [0,T]} \left|\hat{W}_Z^r(t + \frac{1}{r}\hat{R}_B^r(t)) - \hat{W}_Z^r(t)\right| > \epsilon\right) < \epsilon.$$

It then follows from (6.19) that

$$\sup_{t \in [0,T]} \left|\hat{R}_B^r(t) - \hat{W}_B^r(t)\right| \Rightarrow 0 \text{ as } r \to \infty. \tag{6.20}$$

The diffusion scaled version of the equation (6.16) can be written as

$$S^r\left(r^2 t + r\hat{R}_B^r(t), r^2 t + r\hat{R}_B^r(t) + \hat{R}_Z^r(t, v)\right) = v. \tag{6.21}$$

Due to the sharing level K^r, the processing time of a job with size v has bound

$$\hat{R}^r_Z(t) \leq \frac{K^r}{r} v,$$

which is less than $Kv+1$ for all large enough r. By Theorem 5.1, the number of jobs in service $\hat{Z}^r(\cdot)$ converges weakly to $(X^*(\cdot) \wedge K\beta_e)$ as $r \to \infty$. Again, the limiting process is almost surely continuous. So

$$\sup_{t\in[0,T]} \sup_{s\leq K(v+1)/r} \left| \hat{Z}^r(t + \frac{1}{r}\hat{R}^r_B(t) + s) - \hat{Z}^r(t) \right| \Rightarrow 0 \text{ as } r \to \infty.$$

In other words, the number of jobs in service will not oscillate much during the whole service time. Thus

$$\limsup_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \sup_{x\leq kv+1} \left| S^r\Big(r^2 t + r\hat{R}^r_B(t), r^2 t + r\hat{R}^r_B(t) + x\Big)\hat{Z}^r(t) - x \right| > \epsilon \right) < \epsilon.$$

It then follows from (6.21) that, for any $v \geq 0$,

$$\sup_{t\in[0,T]} \left| \hat{R}^r_Z(t, v) - \hat{Z}^r(t)v \right| \Rightarrow 0 \text{ as } r \to \infty. \tag{6.22}$$

By Corollary 5.1, as $r \to \infty$,

$$\hat{Z}^r(\cdot) \Rightarrow (X^*(\cdot) \wedge K),$$

where $X^*(\cdot) = \frac{(W^*(\cdot) - K\beta_e)^+}{\beta} + \frac{W^*(\cdot)\wedge K\beta_e}{\beta_e}$. In fact, this convergence is also a direct application of Theorem 5.1 and the continuous mapping theorem. So the convergence of $\hat{Z}^r(\cdot)$ holds jointly with the convergence in (6.18). In particular,

$$(\hat{W}^r_B(\cdot), \hat{Z}^r(\cdot)) \Rightarrow \big(\beta(X^*(t) - K)^+, (X^*(\cdot) \wedge K)\big) \text{ as } t \to \infty. \tag{6.23}$$

So the joint convergence of $\hat{R}^r_B(\cdot)$ and $\hat{R}^r_Z(\cdot, v)$ follows immediately from (6.20), (6.22) and the above convergence. □

From this proposition, we see that the limiting response times are piece-wise linear and continuous function of the limiting queue size process. It follows from (6.12) and

the continuous mapping theorem that the steady state distribution of the response times are

$$\mathbb{P}\left(R_B^*(\infty) > x\right) = \exp\Big(- \frac{(1+c_s^2)\theta}{c_a^2+c_s^2} K - \frac{2\theta}{c_a^2+c_s^2}\frac{x}{\beta}\Big), \quad x \geq 0, \tag{6.24}$$

$$\mathbb{P}\left(R_Z^*(\infty, v) > x\right) = \exp\Big(- \frac{(1+c_s^2)\theta}{c_a^2+c_s^2}(\frac{x}{v} \wedge K)\Big), \quad x \geq 0, \tag{6.25}$$

$$\mathbb{P}\left(R^*(\infty, v) > x\right) = \begin{cases} \exp\Big(- \frac{(1+c_s^2)\theta}{c_a^2+c_s^2}\frac{x}{v}\Big), & 0 \leq x \leq Kv \\ \exp\Big(- \frac{(1+c_s^2)\theta}{c_a^2+c_s^2} K - \frac{2\theta}{c_a^2+c_s^2}\frac{x-Kv}{\beta}\Big), & x \geq Kv. \end{cases} \tag{6.26}$$

Similar as in Section 6.2.1, we can obtain the result below as a corollary of Theorem 6.1. The difference is that the linear and continuous relationship (6.17) only holds for the heavy traffic limit, not for each rth system, so we can not apply the continuous mapping theorem. However, the coupling inequality (6.7) holds for the response times as well as the measure valued process. (The reason is that if two queues are the same, then the virtual response times will also be the same.) So the following result can be proved following the same approach of proving Theorem 6.1. We omit the proof for brevity.

Corollary 6.2. *Assume* (4.9)–(4.15), (5.5)–(5.9). *For any* $v \geq 0$,

$$(\hat{R}_B^r(\infty), \hat{R}_Z^r(\infty, v)) \Rightarrow (R_B^*(\infty), R_Z^*(\infty, v)),$$

as $r \to \infty$.

6.3 Approximations

In this section, we apply our limit theorems to obtain approximations for the steady-state queue length and response time.

6.3.1 queue size

Since we have validated the heavy traffic steady state approximation, we can use the steady state random variable $X^*(\infty)$ to approximate the steady state of the diffusion

scaled rth system $\hat{X}^r(\infty)$. It follows from (6.12) that

$$\mathbb{E}(X^*(\infty)) = \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{1}{\theta}(1 - d_p^*(\infty)) + \frac{c_a^2 + c_s^2}{2}\frac{1}{\theta}d_p^*(\infty),$$

where $d_p^*(\infty)$ is given in (6.13). According to the heavy traffic conditions (5.8) and (4.12), $\frac{1}{r}$ can be approximately written as $\frac{1-\rho^r}{\rho^r\theta}$ and θK can be approximately written as $\frac{1-\rho^r}{\rho^r}K^r$. So we obtain the following approximation for $\mathbb{E}(X^r(\infty))$:

$$\mathbb{E}(X^r(\infty)) \approx \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{\rho^r}{1-\rho^r}(1 - d_p^r(\infty)) + \frac{c_a^2 + c_s^2}{2}\frac{\rho^r}{1-\rho^r}d_p^r(\infty),$$

where the delay probability $d_p^r(\infty)$ could be taken as $\exp\big(-\frac{1+c_s^2}{c_a^2+c_s^2}\frac{1-\rho^r}{\rho^r}K^r\big)$. Since $\frac{1-\rho^r}{\rho^r} \sim -\ln\rho^r$, we prefer to use the asymptotically equivalent description $d_p^r(\infty) = \big(\rho^r\big)^{\frac{1+c_s^2}{c_a^2+c_s^2}K^r}$.

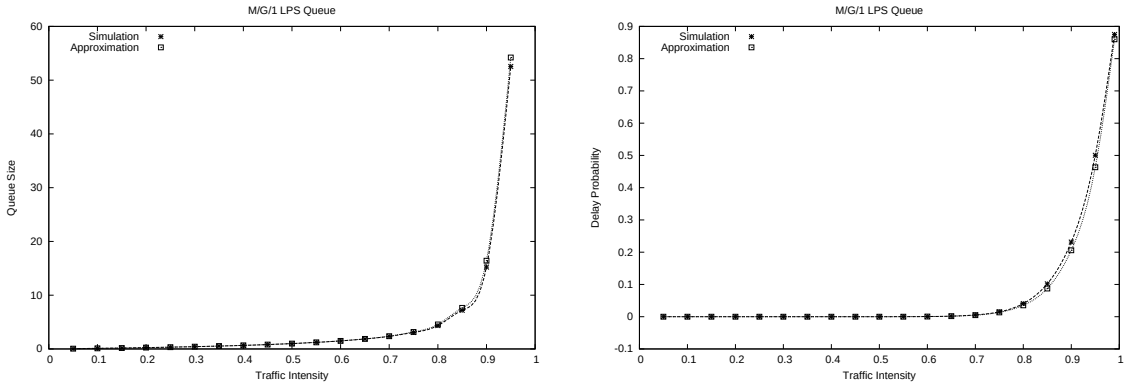

Figure 2: A comparison of the approximation formulas with simulation estimates of steady state response times of the $M/G/1$ LPS Queue.The sharing level $K = 15$, service time distribution is log-normal with $c_s^2 = 9$ and traffic intensities range from 0.05 to 0.95.

In practice, only one system with certain sharing level K and traffic intensity $\rho < 1$ will be given. So we can drop the index r and obtain the following approximation formula:

$$\mathbb{E}[X] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{\rho}{1-\rho}(1 - d_p) + \frac{c_a^2 + c_s^2}{2}\frac{\rho}{1-\rho}d_p, \tag{6.27}$$

$$d_p \approx \rho^{\frac{1+c_s^2}{c_a^2+c_s^2}K}. \tag{6.28}$$

The resulting approximation (6.1) reduces to Kingman's formula for the FIFO queue when the sharing level $K = 1$, and the formula in [22] for the PS queue when the sharing level $K = \infty$. Although the approximation formulas are derived from heavy traffic theorems, they are actually explicit if the arrival process is Poisson and either $K = 1$ or $K = \infty$. In addition, the quality of the approximations is actually reasonable for all traffic intensities, cf. Figure 2.

The approximation formulas are derived in the context of the $G/G/1$ LPS queue, and use the first two moments of inter-arrival time and service time distributions. Table 1 demonstrates the quality of our approximations for various combinations of inter-arrival time and service time distributions. As suggested by the formulas, no matter how we change the combination of distributions, the approximation will be the same as long as the coefficient of variation c_a^2 for inter-arrival time distribution (and c_s^2 for service time) are fixed. This is also reflected by the numerical results in Table 1.

Table 1: $G/G/1$ LPS Queue. The sharing limit $K = 20$, traffic intensity $\rho = 0.9$. The coefficient of variation of inter-arrival time and service time distribution are fixed at $c_a^2 = 4$ and $c_s^2 = 8$ respectively.

Arr. Dist.	Ser. Dist.	$\mathbb{E}[X]$	d_p
HyperExp2p	HyperExp2p	20.5378 ± 0.3148	0.2519 ± 0.0022
	Log-normal	20.6243 ± 0.2753	0.2798 ± 0.0019
	Hyper2star	20.5642 ± 0.1619	0.2066 ± 0.0014
Log-normal	HyperExp2p	19.6028 ± 0.2957	0.2334 ± 0.0020
	Log-normal	19.3615 ± 0.1894	0.2562 ± 0.0017
	Hyper2star	19.5913 ± 0.2435	0.1981 ± 0.0018
Hyper2star	HyperExp2p	20.9429 ± 0.3561	0.2676 ± 0.0027
	Log-normal	21.1600 ± 0.2136	0.2972 ± 0.0015
	Hyper2star	20.8725 ± 0.2505	0.2089 ± 0.0018
Approximation Formulas		20.6474	0.2059

HyperExp2p is the hyper-exponential distribution with 2 phases. A Hyper2star random variable has probability p to be 0 and probability $(1 - p)$ to be an exponential distribution.

6.3.2 Response time

As for the response time, we can use the steady state $R^*_B(\infty)$ and $R^*_Z(\infty, v)$ to approximate the steady state of the diffusion scaled response time of the rth system, i.e. $\hat{R}^r_B(\infty)$ and $\hat{R}^r_Z(\infty, v)$. By (6.24) and (6.25), the following expectations can be easily computed,

$$\mathbb{E}[R^*_B(\infty)] = \frac{c_a^2 + c_s^2}{2}\frac{1}{\theta}p\beta,$$
$$\mathbb{E}[R^*_Z(\infty, v)] = \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{1}{\theta}(1-p)v.$$

Now, we approximate $\frac{1}{r}$ using $\frac{1-\rho^r}{\theta}$. This way of approximating $\frac{1}{r}$ is equivalent in the limit to using $\frac{1-\rho^r}{\rho^r\theta}$ based on the heavy traffic condition (5.8). The main reason for the difference is to make the approximations of waiting time and buffer queue size consistent with Little's law. So we obtain the following approximation formulas for a given system:

$$\mathbb{E}[R_B] \approx \frac{c_a^2 + c_s^2}{2}\frac{1}{1-\rho}d_p\beta, \tag{6.29}$$
$$\mathbb{E}[R_Z(v)] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{1}{1-\rho}(1-d_p)v, \tag{6.30}$$

where d_p is the same as in (6.28).

Table 2 shows the quality of approximations (6.29) and (6.30) for the $M/G/1$ LPS queue with various service time distributions. Again, our approximations use up to the second moment of the service time distribution, so the simulation gives the similar performance for different distributions with the same coefficient of variation c_s^2.

Let R_Z be the unconditional steady state service time, then

$$\mathbb{E}[R_Z] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{1}{1-\rho}(1-d_p)\beta.$$

So we obtain the unconditional response time

$$\mathbb{E}[R] \approx \frac{c_a^2 + c_s^2}{2}\frac{\beta}{1-\rho}d_p + \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{\beta}{1-\rho}(1-d_p). \tag{6.31}$$

Table 2: $M/G/1$ LPS Queue. The sharing limit $K = 30$, traffic intensity $\rho = 0.95$. The coefficient of variation of service time is fixed at $c_s^2 = 19$.

Distribution	$\mathbb{E}[R_B]$	$\mathbb{E}[R_Z(v)/v]$
HyperExp2phase	42.1670 ± 2.0085	15.7579 ± 0.0871
LogNormal	37.2947 ± 1.6338	15.9390 ± 0.0942
Hyper2Star	41.0397 ± 1.6116	15.6039 ± 0.0851
Bimodal	41.8724 ± 1.3550	15.6162 ± 0.0958
Approximations	42.9278	15.7072

Finally, we show in Tables 3 a comparison of all our performance approximations with simulations of the $M/G/1$ and the $G/M/1$ LPS queues. All the numerical results show that the two-moment approximations are reasonable fit, with the exception of log-normal service times (in Table 2). This is in accordance with other numerical studies on the quality of two-moment approximations, see for example [27].

Table 3: $G/M/1$ and $M/G/1$ LPS Queues. The sharing limit $K = 10$, traffic intensity $\rho = 0.9$.

Perf. meas.	$M/G/1$		G/M/1	
	simulation	approx.	simulation	approx.
d_p	0.3437 ± 0.0025	0.3478	0.2436 ± 0.0026	0.2580
$E(X)$	8.2459 ± 0.0569	8.2155	6.8441 ± 0.0555	7.0000
$E(R_B)$	2.6591 ± 0.0466	2.6151	1.8629 ± 0.0426	2.0070
$E(R_Z)$	6.4958 ± 0.0146	6.5132	5.6779 ± 0.0171	5.7708

The service time distribution for $M/G/1$ is Erlang with 2 phases (E_2), mean is 1 and $c_s^2 = 1/2$. The inter-arrival time distribution for $G/M/1$ is E_2 with mean $1/0.9$ and $c_a^2 = 1/(2 \times 0.9)$.

LIMITED PROCESSOR SHARING QUEUES AND MULTI-SERVER QUEUES

VOLUME II

Multi-server Queues

by

Jiheng Zhang

CHAPTER VII

INTRODUCTION

Recently, there has been a great interest in many-server queues with a large number of servers. Such queueing systems have been used extensively to model customer call centers; see for example, survey papers Aksin et al. [1] and Gans et al. [20]. Since a customer can easily hang up after waiting for too long, abandonment is a non-negligible aspect in the study of many-server queues. As pointed out in Garnett et al. [21], customer abandonment is a key factor for call center management. In our study, a customer can leave the system (without getting service) once has been waiting in queue for more than his patience time. Both patience and service times are modeled using random variables. A recent statistical study by Brown et al. [8] suggests that the exponential assumption on service time distribution, in many case, is not valid. In fact, the distribution of service times at call centers may be log-normal in some cases as shown in [8]. This emphasizes the need to look at the many-server model with general service and patience times.

In this paper, we study many-server queues with general patience and service times. The queueing model is denoted by $G/GI/n{+}GI$. The G represents a general stationary arrival process. The first GI indicates that service times come from a sequences of independent and identically distributed (IID) random variables with a general distribution. The n denotes the number of homogeneous servers. There is an unlimited waiting space called buffer, where customers wait and can choose to abandon if their patience times expires before their service starts. Again, the patience times of each customer are IID and with a general distribution (the GI after the '+' sign).

Useful insights can be obtained by considering a many-server queue in limit regimes where the number n of servers increases along with the arrival rate λ^n such that the traffic intensity

$$\rho^n = \frac{\lambda^n}{n\mu} \to \rho \text{ as } n \to \infty,$$

where the μ is the service rate of a single server (in other words, the reciprocal of the mean service time), and $\rho \in [0, \infty)$. Since the abandonment ensures stability, the limit ρ in the above need not to be less than 1. In fact, according to ρ, the limit regimes can be divided into three classes, namely, the *Efficiency-Driven* (ED) regime when $\rho > 1$, the *Quality and Efficiency-Driven* (QED) regime when $\rho = 1$ and the *Quality Driven* (QD) regime when $\rho < 1$. The QED regime is also called the *Halfin-Whitt* regime due to the seminal work Halfin and Whitt [28]. With this motivation, we establish the fluid (also called law of large number) limit for the $G/GI/n{+}GI$ queue in all three regimes.

We show that the fluid model has an equilibrium, which yields approximations for various performance quantities. These fluid approximations work pretty well in the ED and QD regime where ρ is not that close to 1, as demonstrated in the numerical experiments.

One of the challenges in studying many-server queues with general service times (as well as general patience time) is that a simple Markovian analysis is not feasible. In a system where multiple customers are processed at the same time, such as the many-server queue, how to describe the system becomes an important issue. The number of customers in the system does not give much information since they may all have large remaining service times or all have small remaining service times, and this information can affect future evolution of the system. So we choose finite Borel measures on $\mathbb{R}$ to describe the state of the system. At any time $t \geq 0$, instead of recording the total number of customers in service (i.e. the number of busy servers), we record all the remaining service times using measure $\mathcal{Z}(t)$. For any Borel set $C \subseteq (0, \infty)$, $\mathcal{Z}(t)(C)$

indicates the number of customers in server with *remaining service time* belongs to C at that time. Similar idea applies for the remaining patience times. We first introduce the *virtual buffer*, which holds all the customers who have arrived but not yet scheduled to receive service (assuming they are infinitely patient). We record all the remaining patience times for those in the virtual buffer using finite Borel measure $\mathcal{R}(t)$ on $\mathbb{R} = (-\infty, \infty)$. At time $t \geq 0$, $\mathcal{R}(t)(C)$ indicates the number of customers in buffer with *remaining patience time* belongs to the Borel set $C \subseteq \mathbb{R}$. The descriptor $(\mathcal{R}(\cdot), \mathcal{Z}(\cdot))$ contains very rich information, almost all information about the system can be recovered from it. Note that a customer with negative remaining patience time has already abandoned. So the actual number of customers in the buffer is

$$Q(t) = \mathcal{R}(t)((0, \infty)) \text{ for all } t \geq 0.$$

More details will be discussed when we rigorously introduce the mathematical model in Section 8. In the literature, another descriptor that keeps track of the ages of customers in service and the ages of customers in waiting have been used, e.g. [35, 56]; The age processes have the advantage of being observable, without requiring future information, though their analysis is often more complicated. Both age and residual descriptions of the system often results in the same steady state insights. In this part of the thesis, we focus on residual processes only.

The framework of using measure-valued process has been successfully applied to study models where multiple customers are processed at the same time. Existing works include Gromoll and Kruk [24], Gromoll, Puha and Williams [25] and Gromoll, Robert and Zwart [26], to name a few. Most of these works are on the processor sharing queue and related models where there is no waiting buffer. Recently, Zhang, Dai and Zwart [63, 62] apply the measure-valued process to study the limited processor sharing queue, where only limited number of customers can be served at any time with extra customers waiting in a buffer; see also the first part of the thesis. Many techniques in the present part of the thesis closely follows from those developed in

[63]. There has been a huge literature on many-server queue and related models since the seminal work by Halfin and Whitt [28]. But there are not many successes with the case where the service time distribution is allowed to be non-exponential. One exception is the work of Reed [48], in which fluid and diffusion limits of the customer-count process of many server queues (without abandonment) are established where few assumptions beyond a first moment are placed on the service time distribution. Later, Puhalskii and Reed [46] extend the aforementioned results to allow noncritical loading, generally distributed service times, and general initial conditions. Jelenković et al. [32] study the many-server queue with deterministic service times; Garmarnik and Momčilović [18] study the model with lattice-valued service times; Puhalskii and Reiman [47] study the model with phase-type service time distributions. Mandelbaum and Momčilović [39] study the virtual waiting time processes, and Kaspi and Ramanan [36] study the fluid limit of measure-valued processes for many-server queues with general service times. For the many-server queue with abandonment, a version of the fluid model have been established as a conjecture in Whitt [55], where a lot of insight was demonstrated, which help greatly in our work. Recently, Kang and Ramanan also worked on the same topic and summarized their result in the technical report [35]. Although we focus on the same topic, our work uses different methodology from that in [35] and requires less assumptions on the service time distribution. Our approach mainly based on tracking the "residual" processes, while [35] tracks the "age" processes for studying the queueing model. In our work, the only assumption on the service time distribution is continuity, while the service time distribution in [35] is required to have a density and the hazard rate function must be bounded. Also, in our analysis, we use a fluid model which is defined in a simpler way. This facilitates the analysis. In addition, we verify in the appendix of this thesis (c.f. Appendix F) that our fluid model is consistent with the special case where both service and patience times are exponentially distributed, as established in Whitt [55].

Additional works on many-server queues with abandonment includes Dai, He and Tezcan [13] for phase-type service time distributions and exponential patience time distribution; Zeltyn and Mandelbaum [58] for exponential service time distribution and general patience time distributions; Mandelbaum and Momčilović [40] for both general service time distribution and general patience time distribution.

This part of the thesis is organized as follows: We begin in Chapter 8 by formulating the mathematical model of the $G/GI/n{+}GI$ queue. The dynamics of the system are clearly described by modeling with measure valued processes; see (8.4) and (8.5). In Chapter 9, we explore the fluid model and its properties. Chapter 10 is devoted to establishing the convergence of stochastic process which includes the proof of pre-compactness and the characterization of the limit as the fluid model solution. This part of thesis is summarized in the technical report [61] .

CHAPTER VIII

STOCHASTIC MODEL

In this chapter, we first describe the $G/GI/n{+}GI$ queueing system and then dscribe a pair of measure-valued processes that capture the dynamics of the system. On the service side, similar as the LPS queue, we use a measure valued descriptor to keep track of every unfinished work. The new idea in this study is to use measure valued descriptor to keep track of residual patient time of each waiting customer. In the following, we first introduce necessary notations and then show how to use the measure-valued descriptors to describe the dynamics of the system.

There are n identical servers in the system. Customers arrive according to a general stationary arrival process (the initial G) with arrival rate λ. Let a_i denote the arrive time of the ith arriving customer, $i = 1, 2, \cdots$. Arriving customer enters service immediately upon arrival if there is a server available. If all n servers are busy, the arriving customer waits in a buffer, which has infinite capacity. Customers are served in the order of their arrival by the first available server. Waiting customers may also elect to abandon. We assume that each customer has a random patience time. A customer will abandon immediately when his waiting time in the buffer exceeds his patience time. Once a customer starts his service, the customer remains until the service is completed. There are no retrials; abandoning customers leave without affecting future arrivals.

The two GIs in the notation mean that the service times and patience times come from two independent sequences of iid random variables; these two sequences are assumed to be independent of the arrival process. Let u_i and v_i denote the patience and service time of the ith arriving customer, $i = 1, 2, \cdots$. In many applications such

as telephone call centers, customers cannot see the queue (the case of invisible queues, c.f. [41]), thus do not know the experience of other customers. In such a case, it is natural to assume that patience times are iid. Denote $F(\cdot)$ and $G(\cdot)$ the distributions for the patience and service times, respectively.

To describe the system using measure-valued process, we first introduce the notion of *virtual buffer*. The virtual buffer holds all customers in the real buffer and some of the abandoned customers. An abandoned customer continues to wait in the virtual buffer when he first abandons until it were his turn for service had he not abandoned. At this time, he leaves the virtual buffer. At any time $t \geq 0$, $\mathcal{R}(t)$ denotes a measure in $\mathbf{M}$ such that $\mathcal{R}(t)(C)$ is the number of customers in the virtual buffer with remaining patience time in $C \in \mathscr{B}(\mathbb{R})$. Please note that this way of modeling requires $\mathcal{R}(\cdot)$ to be a measure on $\mathbb{R}$, not just $\mathbb{R}_+$. It is clear that

$$Q(t) = \mathcal{R}(t)(\mathbb{R}_+) \text{ and } R(t) = \mathcal{R}(t)(\mathbb{R}) \tag{8.1}$$

represent the number of customers waiting in the real buffer and number of customers in the virtual buffer, respectively.

We also use a measure to describe the server. At any time $t \geq 0$, $\mathcal{Z}(t)$ denotes a measure in $\mathbf{M}_+$ such that $\mathcal{Z}(t)(C)$ is the number of customers in service with remaining service time in $C \in \mathscr{B}(\mathbb{R}_+)$. Different from the virtual buffer, the servers only hold customers with positive remaining service times, so we only care about the subsets in $\mathbb{R}_+$. The quantity

$$Z(t) = \mathcal{Z}(t)(\mathbb{R}_+), \tag{8.2}$$

represents the number of customers in service at any time $t \geq 0$.

The measure-valued (taking value in $\mathbf{M} \times \mathbf{M}_2$) stochastic process $(\mathcal{R}(\cdot), \mathcal{Z}(\cdot))$ serves as the descriptor for the $G/GI/n{+}GI$ queueing model. Before we use this to describe the dynamics of the system, let us first talk about the initial condition, since the system is allowed to be non-empty initially. The initial state specifies $R(0)$, the

number of customers in the virtual buffer as well as their remaining patience times u_i and service times v_i, $i = 1 - R(0), 2 - R(0), \cdots, 0$. The initial state also specifies $Z(0)$, the number of customers in service as well as their remaining service times v_i, $i = 1 - R(0) - Z(0), \cdots, -R(0)$. Briefly, the initial customers are given negative index, in order not to conflict with the index of arriving customers. Those initial customers in the buffer are also assumed to have i.i.d. service times with distribution $G(\cdot)$. For each $t \geq 0$, denote $E(t)$ the number of customers that has arrived during time interval $(0, t]$. Arriving customers are indexed by $1, 2, \cdots$ according to the order of their arrival. By this way of indexing customers, it is clear that the index of the first customer in the virtual buffer at time $t \geq 0$ is $B(t) + 1$, where

$$B(t) = E(t) - R(t). \tag{8.3}$$

Denote w_i the waiting time of the ith customers; then $\tau_i = a_i + w_i$ is the time that the ith job starts *service* for all $i \geq 1 - R(0)$. When $i < 0$, a_i may be a negative number indicating how long the ith customer had been there by time 0. We will impose some conditions on a_i's with $i < 0$ later on. Let δ_x and $\delta_{(x,y)}$ denote the Dirac point measure at $x \in \mathbb{R}$ and $(x, y) \in \mathbb{R}^2$, respectively. Denote $C + x = \{c + x : x \in C\}$ for any subset $C \subset \mathbb{R}$ and $C_x = (x, \infty)$. For any subsets $C, C' \subset \mathbb{R}$, let $C \times C'$ denote the Cartesian product. Using the Dirac measure and the above introduced notations, the evolution of the system can be captured by the following *stochastic dynamic equations*:

$$\mathcal{R}(t)(C) = \sum_{i=1+B(t)}^{E(t)} \delta_{u_i}(C + t - a_i), \quad \text{for all } C \in \mathscr{B}(\mathbb{R}), \tag{8.4}$$

$$\begin{aligned} \mathcal{Z}(t)(C) = & \sum_{i=1-R(0)-Z(0)}^{-R(0)} \delta_{v_i}(C + t) \\ & + \sum_{i=1-R(0)}^{B(t)} \delta_{(u_i,v_i)}(C_0 + \tau_i - a_i) \times (C + t - \tau_i), \end{aligned} \quad \text{for all } C \in \mathscr{B}(\mathbb{R}_+), \tag{8.5}$$

at any time $t \geq 0$. Denote the total number of customers in the system by

$$X(t) = Q(t) + Z(t) \quad \text{for all } t \geq 0. \tag{8.6}$$

The following *policy constraints* must be satisfied at any time $t \geq 0$,

$$Q(t) = (X(t) - n)^+, \tag{8.7}$$

$$Z(t) = (X(t) \wedge n), \tag{8.8}$$

where n, as introduced above, denotes the number of servers in the system.

CHAPTER IX

FLUID MODEL AND ITS PROPERTIES

To study the stochastic model, we introduce a determinisitic fluid model. This fluid model will be shown later to be the approximation of certain scaled stochastic processes. Similar to the fluid model for the LPS queue, the major challenge from the mathematical point view is the function equation (9.21), which is derived from the fluid model.

9.1 Fluid Model

Denote the means of the marginal distributions F and G to be $1/\alpha$ and $1/\mu$, respectively. To simplify notations, let $F^c(\cdot)$ denote the complement of the probability distribution function $F(\cdot)$, i.e. $F^c(x) = 1 - F(x)$ for all $x \in \mathbb{R}$; $G^c(\cdot)$ is defined similarly. We introduce the following *fluid dynamic equations*:

$$\bar{\mathcal{R}}(t)(C_x) = \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F^c(x+t-s)ds, \quad t \geq 0, \quad x \in \mathbb{R}, \tag{9.1}$$

$$\bar{\mathcal{Z}}(t)(C_x) = \bar{\mathcal{Z}}(0)(C_x + t) + \int_0^t F^c(\frac{\bar{R}(s)}{\lambda})G^c(x+t-s)d\bar{B}(s), \quad t \geq 0, \quad x \in \mathbb{R}_+, \tag{9.2}$$

where $C_x = (x, \infty)$ and $\bar{B}(s) = \lambda s - \bar{R}(s)$. Here, all the time dependent quanities are assumed to be right continuous on $[0, \infty)$ and to have left limits in $(0, \infty)$; furthermore, the integral $\int_0^t g(s)\, d\bar{B}(s)$ is interpreted as the Lebesgue-Stieltjes integral on the interval $(0, t]$. The quantities $\bar{R}(\cdot)$, $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$ and $\bar{X}(\cdot)$ are defined in the same way as their stochastic counterparts in (8.1), (8.2) and (8.6). The following policy

constraints must be satisfied

$$\bar{Q}(t) = (\bar{X}(t) - 1)^+, \tag{9.3}$$

$$\bar{Z}(t) = (\bar{X}(t) \wedge 1). \tag{9.4}$$

The fluid dynamic equations (9.1) and (9.2) and the policy constraints (9.3) and (9.4) define a *fluid model*, which is denoted by (λ, H).

Denote $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0) = (\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0))$ to be the initial condition of the fluid model. For the convenience of notations, also denote $\bar{Q}_0 = \bar{Q}(0)$, $\bar{Z}_0 = \bar{Z}(0)$ and $\bar{X}_0 = \bar{Q}_0 + \bar{Z}_0$. We need to assume that the initial condition satisfies the dynamic equations and the policy constraints, i.e.

$$\bar{\mathcal{R}}_0(C_x) = \lambda \int_0^{\frac{\bar{R}_0}{\lambda}} F^c(x+s)ds, \quad x \in \mathbb{R}, \tag{9.5}$$

$$\bar{Q}_0 = (\bar{X}_0 - 1)^+, \tag{9.6}$$

$$\bar{Z}_0 = (\bar{X}_0 \wedge 1). \tag{9.7}$$

We also require that

$$\bar{\mathcal{Z}}_0(\{0\}) = 0, \tag{9.8}$$

which means that nobody with remaining service time 0 stays in the server. We call any element $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0) \in \mathbf{M} \times \mathbf{M}_2$ a *valid* initial condition if it satisfies (9.5)–(9.8).

We call $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M} \times \mathbf{M}_2)$ a solution to the fluid model (λ, H) with a valid initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$ if it satisfies the fluid dynamic equations (9.1) and (9.2) and the policy constraints (9.3) and (9.4).

We need to assume that the patience time distribution $F(\cdot)$ has density $f(\cdot)$. Let

$$M_F = \inf\{x \geq 0 : F(x) = 1\}. \tag{9.9}$$

By the right continuity of F, it is clear that $F(x) < 1$ for all $x < M_F$ and $F(x) = 1$ for all $x \geq M_F$. Define the hazard rate $h_F(\cdot)$ of the distribution $F(\cdot)$ by

$$h_F(x) = \begin{cases} \frac{f(x)}{1-F(x)} & x < M_F, \\ 0 & x \geq M_F. \end{cases}$$

Theorem 9.1 (Existence and Uniqueness)**.** *Assume that the service time distribution* $G(\cdot)$ *satisfies*

$$G(\cdot) \text{ is continuous,} \tag{9.10}$$

$$0 < \mu < \infty, \tag{9.11}$$

and the patience time distribution $F(\cdot)$ *has a density* $f(\cdot)$ *such that the hazard rate satisfies*

$$0 < \alpha < \infty, \tag{9.12}$$

$$\sup_{x \in [0,\infty)} h_F(x) < \infty. \tag{9.13}$$

There exists a unique solution to the fluid model (λ, H) *for any valid initial condition* $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

The above theorem provides the foundation to further study the fluid model. A key property is that the fluid model has an equilibrium state. An equilibrium state is defined as the following:

Definition 9.1. *An element* $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty) \in \mathbf{M} \times \mathbf{M}_2$ *is called an* equilibrium state *for the fluid model* (λ, H) *if the solution to the fluid model with initial condition* $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ *satisfies*

$$(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty) \quad \textit{for all } t \geq 0.$$

This definition says that if a fluid model solution starts from an equilibrium state, it will never change in the future. To present the result about equilibrium state, we need to introduce some notation. For the service time distribution function $G(\cdot)$ on $\mathbb{R}_+$, the *equilibrium* distribution associated with G is given by

$$G_e(x) = \mu \int_0^x G^c(y)dy, \quad \text{for all } x \geq 0.$$

Theorem 9.2. *Assume the conditions in Theorem 9.1. The state* $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ *is an equilibrium state of the fluid model* (λ, H) *if and only if it satisfies*

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_0^w F^c(x+s)ds, \quad x \in \mathbb{R}, \tag{9.14}$$

$$\bar{\mathcal{Z}}_\infty(C_x) = \min(\rho, 1)\,[1 - G_e(x)], \quad x \in \mathbb{R}_+, \tag{9.15}$$

where w *is a solution to the equation*

$$F(w) = \max\left(\frac{\rho - 1}{\rho}, 0\right). \tag{9.16}$$

Remark 9.1. *If equation* (9.16) *has multiple solutions, then the equilibrium is not unique (any solution* w *gives an equilibrium). If the equation has a unique solution (for example when* $F(\cdot)$ *is strictly increasing), then the equilibrium state is unique.*

The quantity w, when it is unique, is interpreted to be the *offered* waiting time for an arriving customer. If his patience time exceeds w, he will not abandon. Thus, the probabilty of his abandonment is given by $F(w)$, which is equal to $(\rho-1)/\rho$ when $\rho > 1$; the latter quantity is the fraction of traffic that has to be discarded due to the overloading. From (9.14), $\bar{\mathcal{R}}_\infty(C_x) = \lambda w$ for $x \le -w$. Thus, the average number of customers in the virtual buffer is

$$\bar{R}_\infty = \bar{\mathcal{R}}_\infty(\mathbb{R}) = \lambda w,$$

which is consistent with Little's law. From (9.15), the average number of busy servers is

$$\bar{Z}_\infty = \bar{\mathcal{Z}}_\infty(\mathbb{R}_+) = \min(\rho, 1),$$

which is intuitively clear. These observations and interpreprations were first made by Whitt; these fluid model based performance measures have been used to approximate performance measures in the stochastic system in [56].

9.2 Existence and Uniqueness of Fluid Model Solutions

It follows from (9.1) that

$$\begin{aligned}\bar{Q}(t) = \bar{\mathcal{R}}(t)(C_0) &= \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F^c(t-s)ds \\ &= \lambda \int_0^{\frac{\bar{R}(t)}{\lambda}} F^c(s)ds.\end{aligned}$$

Let $F_e(\cdot)$ denote the equilibrium distribution associated with distribution $F(\cdot)$ (defined in the same way as $G_e(\cdot)$). So we get

$$\bar{Q}(t) = \frac{\lambda}{\alpha} F_e(\frac{\bar{R}(t)}{\lambda}). \tag{9.17}$$

Recall the support M_F defined in (9.9). It is clear that $F_e(x)$ is strictly monotone for $x \in [0, M_F)$. Thus, $F_e^{-1}(y)$ is well defined for each $y \in [0, 1)$. When M_F is finite, $F_e(M_F) = 1$, and we we take $F_e^{-1}(1) = M_F$. Thus, (9.17) implies that $\alpha\bar{Q}(t)/\lambda \leq 1$ for all $t \geq 0$, and

$$\bar{R}(t)/\lambda = F_e^{-1}(\alpha\bar{Q}(t)/\lambda) \quad \text{ for } t \geq 0. \tag{9.18}$$

It follows from (9.2) that

$$\begin{aligned}\bar{Z}(t) = \bar{\mathcal{Z}}(t)(C_0) = \bar{\mathcal{Z}}_0(C_0 + t) &+ \lambda \int_0^t F^c(\frac{\bar{R}(s)}{\lambda}) G^c(t-s)ds \\ &- \int_0^t F^c(\frac{\bar{R}(s)}{\lambda}) G^c(t-s)d\bar{R}(s).\end{aligned}$$

Note that by (9.17), $d\bar{Q}(s) = F^c(\frac{\bar{R}(s)}{\lambda})d\bar{R}(s)$. So

$$\bar{Z}(t) = \bar{\mathcal{Z}}_0(C_0 + t) + \frac{\lambda}{\mu}\int_0^t F^c(\frac{\bar{R}(s)}{\lambda})dG_e(t-s) - \int_0^t G^c(t-s)d\bar{Q}(s).$$

So we get

$$\begin{aligned}\bar{Z}(t) = \bar{\mathcal{Z}}_0(C_t) &+ \frac{\lambda}{\mu}\int_0^t F^c(\frac{\bar{R}(t-s)}{\lambda})dG_e(s) \\ &- \bar{Q}(t)G^c(0) + \bar{Q}(0)G^c(t) + \int_0^t \bar{Q}(t-s)dG(s).\end{aligned} \tag{9.19}$$

We wish to represent the term $F^c(\frac{\bar{R}(t-s)}{\lambda})$ using $\bar{Q}(\cdot)$. By (9.18) we have that

$$F^c(\frac{\bar{R}(t)}{\lambda}) = F^c\left(F_e^{-1}(\frac{\alpha}{\lambda}\bar{Q}(t))\right), \tag{9.20}$$

for all $t \geq 0$. Note that $G^c(0) = 1$ by assumption (9.10). Combining (9.3), (9.4), (9.19), and (9.20), we obtain

$$\begin{aligned}\bar{X}(t) = \bar{\mathcal{Z}}_0(C_t) + \bar{Q}_0 G^c(t) \\ + \frac{\lambda}{\mu}\int_0^t F^c\Big(F_e^{-1}(\frac{\alpha}{\lambda}(\bar{X}(t-s)-1)^+)\Big)dG_e(s) \\ + \int_0^t (\bar{X}(t-s)-1)^+ dG(s).\end{aligned}$$

For technical reason, we need to extend the domain of function $F_e^{-1}(\cdot)$ from $[0,1]$ to $[0,\infty)$ by defining $F_e^{-1}(x) = M_F$ for all $x \geq 1$. To simplify the notation, denote $\zeta_0(\cdot) = \bar{\mathcal{Z}}_0(C_0 + \cdot) + \bar{Q}_0 G^c(\cdot)$ and $H(x) = F^c(F_e^{-1}(\frac{\alpha}{\lambda}x))$. It follows that

$$\bar{X}(t) = \zeta_0(t) + \rho\int_0^t H\big((\bar{X}(t-s)-1)^+\big)dG_e(s) + \int_0^t (\bar{X}(t-s)-1)^+ dG(s). \quad (9.21)$$

Please note that $\zeta_0(\cdot)$ depends only on the initial condition and $H(\cdot)$ is a function defined by the arrival rate λ and the patience time distribution $F(\cdot)$. The equation (9.21) will serve as a key to the analysis of the fluid model.

We establish the existence and uniqueness (Theorem 9.1) of the fluid model solution. This provides the foundation on which we can further explore the properties of the fluid model.

Existence Although we have established the existence of the solution to (9.21) in Lemma E.1, some technical issues still exist in establishing the existence of fluid model solution. So we can not construct a fluid model solution via the similar method as in Chapter 3 for the LPS queues. In this thesis, we establish the existence of the fluid model solution via fluid limits.

In Lemma 10.6, we show that every fluid limit satisfies the fluid model equations (9.1) and (9.2) and the constraints (9.3) and (9.4). Thus this already establish the existence.

Uniqueness Suppose there is another solution to the fluid model (λ, H) with initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$, denoted by $(\bar{\mathcal{R}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$. Similarly, denote

$$\bar{R}^\dagger(t) = \bar{\mathcal{R}}^\dagger(\mathbb{R}),$$
$$\bar{Z}^\dagger(t) = \bar{\mathcal{Z}}^\dagger(\mathbb{R}_+),$$

for all $t \geq 0$. It must satisfy the fluid dynamic equations (9.1) and (9.2) and constraints (9.3) and (9.4). For all $t \geq 0$, let

$$\bar{Q}^\dagger(t) = \frac{\lambda}{\alpha} \bar{F}_e(\frac{\bar{R}^\dagger(t)}{\lambda}).$$

According to the algebra at the beginning of Section 3.2, $\bar{X}^\dagger(\cdot)$ must satisfy equation (9.21). By the uniqueness of the solution to the equation (9.21)

$$\bar{X}^\dagger(t) = \bar{X}(t) \quad \text{for all } t \geq 0.$$

By (9.3) and (9.18), $\bar{R}^\dagger(t) = \bar{R}(t)$. By the dynamic equations (9.1) and (9.2), we must have that

$$(\bar{\mathcal{R}}^\dagger(t), \bar{\mathcal{Z}}^\dagger(t)) = (\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)) \quad \text{for all } t \geq 0.$$

This proves uniqueness.

9.3 Equilibrium State of the Fluid Model Solution

In this section, we first intuitively explain what an equilibrium should be. Then we rigorously prove it in Theorem 9.2. To provide some intuition, note that in the equilibrium, by equation (9.1), one should have

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_0^{\bar{R}_\infty/\lambda} F^c(x+s)ds,$$

for the buffer. This immediately implies that

$$\bar{\mathcal{R}}_\infty(C_x) = \frac{\lambda}{\alpha}[F_e(x + \frac{\bar{R}_\infty}{\lambda}) - F_e(x)].$$

So the rate at which customers leave the buffer because of impatience is:

$$\lim_{x\to 0}\frac{\bar{\mathcal{R}}_\infty(C_0)-\bar{\mathcal{R}}_\infty(C_x)}{x}=\lambda F(\frac{\bar{R}_\infty}{\lambda}).$$

In the equilibrium, intuitively, the number of customers in service should not change and the distribution for the remaining service time should be the equilibrium distribution $G_e(\cdot)$, i.e.

$$\bar{\mathcal{Z}}_\infty(C_x)=\bar{Z}_\infty[1-G_e(x)].$$

The rate at which customers depart from the server is:

$$\lim_{x\to 0}\frac{\bar{\mathcal{Z}}_\infty(C_0)-\bar{\mathcal{Z}}_\infty(C_x)}{x}=\bar{Z}_\infty\mu.$$

The arrival rate must be equal to the summation of the departure rate from server (due to service completion) and the one from buffer (due to abandonment), i.e.

$$\lambda=\lambda F(\frac{\bar{R}_\infty}{\lambda})+\bar{Z}_\infty\mu. \tag{9.22}$$

It follows directly from (9.17) that

$$\bar{Q}_\infty=\frac{\lambda}{\alpha}F_e(\frac{\bar{R}_\infty}{\lambda}). \tag{9.23}$$

If $\bar{R}_\infty>0$, then according to (9.23) we have $\bar{Q}_\infty>0$. Thus $\bar{Z}_\infty=1$ according to policy constraints. By (9.22), $\rho>1$ and $\frac{\bar{R}_\infty}{\lambda}$ is a solution to the equation $F(w)=\frac{\rho-1}{\rho}$. If $\bar{R}_\infty=0$, then according to (9.22) we have $\rho=\bar{Z}_\infty\leq 1$. In summary, we have that

$$\bar{Q}_\infty=\frac{\lambda}{\alpha}F_e(w),$$
$$\bar{Z}_\infty=\min(\rho,1),$$

where w is a solution to the equation $F(w)=\max(\frac{\rho-1}{\rho},0)$. This is consistent with the one in [56], which is derived from a conjecture of a fluid model. Now, we rigorously prove this result.

Proof of Theorem 9.2. If $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is an equilibrium state, then according to the definition, it must satisfies

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_{t-\frac{\bar{R}_\infty}{\lambda}}^{t} F^c(x+t-s)ds, \quad t \geq 0, \tag{9.24}$$

$$\bar{\mathcal{Z}}_\infty(C_x) = \bar{\mathcal{Z}}_\infty(C_x+t) + \int_0^t F^c(\frac{\bar{R}_\infty}{\lambda})G^c(x+t-s)d\lambda s, \quad t \geq 0. \tag{9.25}$$

It follows from (9.25) that

$$\begin{aligned}\bar{\mathcal{Z}}_\infty(C_x) - \bar{\mathcal{Z}}_\infty(C_x+t) &= \rho F^c(\frac{\bar{R}_\infty}{\lambda})\mu \int_0^t G^c(x+t-s)ds \\ &= \rho F^c(\frac{\bar{R}_\infty}{\lambda})[G_e(x+t) - G_e(x)], \quad t \geq 0.\end{aligned}$$

Taking $t \to \infty$, one has

$$\bar{\mathcal{Z}}_\infty(C_x) = \rho F^c(\frac{\bar{R}_\infty}{\lambda})[1 - G_e(x)]. \tag{9.26}$$

Thus $\bar{Z}_\infty = \rho F^c(\frac{\bar{R}_\infty}{\lambda})$. According to (9.17), we have that

$$\bar{Q}_\infty = \frac{\lambda}{\alpha} F_e(\frac{\bar{R}_\infty}{\lambda}).$$

First assume that $\bar{R}_\infty > 0$. Then $\bar{Q}_\infty > 0$, and thus $\bar{Z}_\infty = 1$ by the policy constraints (9.3) and (9.4). Therefore, $\rho F^c(\frac{\bar{R}_\infty}{\lambda}) = 1$, which implies that $F(\frac{\bar{R}_\infty}{\lambda}) = \frac{\rho-1}{\rho}$ and $\rho > 1$. Now assume that $\bar{R}_\infty = 0$. Then $\bar{Z}_\infty = \rho$, which must be less than or equal to 1 by the policy constraints. Summarizing the cases where $\rho > 1$ and $\rho \leq 1$, we have that the equilibrium state must satisfy (9.14)–(9.16).

If a state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ satisfies (9.14)–(9.16), then let

$$(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty),$$

for all $t \geq 0$. If $\rho \leq 1$, then $\bar{\mathcal{R}}(\cdot) \equiv \mathbf{0}$ and $\bar{Z}(\cdot) \equiv \rho$; if $\rho > 1$, then $\bar{R}(\cdot) \equiv \lambda w$ and $\bar{Z}(\cdot) \equiv 1$, where w is a solution to equation (9.16). It is easy to check that $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is a fluid model solution in both cases. So by definition, the state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is a equilibrium state. □

CHAPTER X

FLUID APPROXIMATION OF THE STOCHASTIC MODELS

We consider a sequence of queueing systems indexed by the number of servers n, where $n \to \infty$. Each model is defined in the same way as in Chapter 8. The arrival rate of each model is assumed be to proportional to n. To distinguish models with different indices, quantities of the nth model are accompanied by superscript n. Each model may be defined on a different probability space $(\Omega^n, \mathcal{F}^n, \mathbb{P}^n)$. Our results concern the asymptotic behavior of the descriptor under the *fluid* scaling, which is defined by

$$\bar{\mathcal{R}}^n(t) = \frac{1}{n}\mathcal{R}^n(t), \quad \bar{\mathcal{Z}}^n(t) = \frac{1}{n}\mathcal{Z}^n(t), \tag{10.1}$$

for all $t \geq 0$. The fluid scaling for the arrival process $E^n(\cdot)$ is defined in the same way, i.e.

$$\bar{E}^n(t) = \frac{1}{n}E^n(t),$$

for all $t \geq 0$. We assume that

$$\bar{E}^n(\cdot) \Rightarrow \lambda \cdot \quad \text{as } n \to \infty. \tag{10.2}$$

Since the limit is deterministic, the convergence in distribution in (10.2) is equivalent to convergence in probability; namely, for each $T > 0$ and each $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}^n\Big(\sup_{0\leq t\leq T} |\bar{E}^n(t) - \lambda t| > \epsilon\Big) = 0.$$

Denote ν_F^n and ν_G^n the probability measures corresponding to the patience time distribution F^n and the service time distribution G^n, respectively. Assume that as $n \to \infty$,

$$\nu_F^n \to \nu_F, \quad \nu_G^n \to \nu_G, \tag{10.3}$$

where ν_F and ν_G are some probability measures with associated distribution functions F and G. Also, the following initial condition will be assumed:

$$(\bar{\mathcal{R}}^n(0), \bar{\mathcal{Z}}^n(0)) \Rightarrow (\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0) \quad \text{as } n \to \infty, \tag{10.4}$$

where, almost surely, $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$ is a valid initial condition and

$$\bar{\mathcal{R}}_0 \text{ and } \bar{\mathcal{Z}}_0 \text{ has no atoms.} \tag{10.5}$$

Theorem 10.1. *In addition to the assumptions* (9.10)–(9.13) *in Theorem 9.1, if the sequence of multi-server queues satisfies* (10.2)–(10.5)*, then*

$$(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)) \Rightarrow (\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \quad \textit{as } n \to \infty,$$

where, almost surely, $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *is the unique solution to the fluid model* (λ, H) *with initial condition* $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

10.1 Precompactness

We first establish the following precompactness for the sequence of fluid scaled stochastic processes $\{(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot))\}$.

Theorem 10.2. *Assume* (4.8)–(4.15)*. The sequence of the fluid scaled stochastic processes* $\{(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot))\}_{N\in\mathbb{N}}$ *is precompact as* $n \to \infty$*; namely, for each subsequence* $\{(\bar{\mathcal{R}}^{n_k}(\cdot), \bar{\mathcal{Z}}^{n_k}(\cdot))\}_{n_k}$ *with* $n_k \to \infty$*, there exists a further subsequence* $\{(\bar{\mathcal{R}}^{n_{k_j}}(\cdot), \bar{\mathcal{Z}}^{n_{k_j}}(\cdot))\}_{n_{k_j}}$ *such that*

$$(\bar{\mathcal{R}}^{n_{k_j}}(\cdot), \bar{\mathcal{Z}}^{n_{k_j}}(\cdot)) \Rightarrow (\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \quad \textit{as } j \to \infty,$$

for some $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M} \times \mathbf{M}_+)$.

The remaining of this section is devoted to proving the above theorem. By Theorem 3.7.2 in [17], it suffices to verify (a) the compact containment property, Lemma 10.2 and (b) the oscillation bound, Lemma 10.5 below.

Similar to (2.3), let

$$B^n(t) = E^n(t) - R^n(t). \tag{10.6}$$

It follows from (8.4) and (8.5) that the dynamics for the fluid scaled processes can be written as

$$\bar{\mathcal{R}}^n(t)(C) = \frac{1}{n}\sum_{i=B^n(t)+1}^{E^n(t)} \delta_{u_i^n}(C + t - a_i^n), \quad \text{for all } C \in \mathscr{B}(\mathbb{R}), \tag{10.7}$$

$$\begin{aligned}\bar{\mathcal{Z}}^n(t)(C) &= \bar{\mathcal{Z}}^n(s)(C + t - s) \\ &\quad + \frac{1}{n}\sum_{i=B^n(s)+1}^{B^n(t)} \delta_{(u_i^n, v_i^n)}(C_0 + \tau_i^n - a_i^n) \times (C + t - \tau_i^n),\end{aligned} \quad \text{for all } C \in \mathscr{B}(\mathbb{R}_+), \tag{10.8}$$

for all $0 \le s \le t$.

10.1.1 Some Preliminary Estimates

To better structure the presentation, we first present some preliminary results, which are built on the Glivenko-Catelli estimates inn Appendix D.

For each n, let $\{u_i^n\}_{i\in\mathbb{Z}}$ be a sequence of i.i.d. random variables with probability measure $\nu_F^n(\cdot)$, let $\{u_i^n\}_{i\in\mathbb{Z}}$ be a sequence of i.i.d. random variables with probability measure $\nu_G^n(\cdot)$. For any $n, m \in \mathbb{Z}$ and $l \in \mathbb{R}_+$, define

$$\bar{\mathcal{L}}_F^n(m, l) = \frac{1}{n}\sum_{i=m+1}^{m+\lfloor nl\rfloor} \delta_{u_i^n}, \tag{10.9}$$

$$\bar{\mathcal{L}}_G^n(m, l) = \frac{1}{n}\sum_{i=m+1}^{m+\lfloor nl\rfloor} \delta_{v_i^n}, \tag{10.10}$$

$$\bar{\mathcal{L}}_{F,G}^n(m, l) = \frac{1}{n}\sum_{i=m+1}^{m+\lfloor nl\rfloor} \delta_{(u_i^n, v_i^n)}, \tag{10.11}$$

where δ_x denotes the Dirac measure of point x on $\mathbb{R}$ and $\delta_{(x,y)}$ denotes the Dirac measure of point (x, y) on $\mathbb{R} \times \mathbb{R}$. So $\bar{\mathcal{L}}_F^n(m, l)$ and $\bar{\mathcal{L}}_G^n(m, l)$ are measures on $\mathbb{R}$ and $\bar{\mathcal{L}}_{F,G}^n(m, l)$ is a measure on $\mathbb{R} \times \mathbb{R}$.

Denote $C_x = (x, \infty)$, for all $x \in \mathbb{R}$. We define two classes of testing functions by

$$\mathscr{V} = \left\{1_{C_x}(\cdot) : x \in \mathbb{R}\right\},$$
$$\mathscr{V}_2 = \left\{1_{C_x \times C_y}(\cdot, \cdot) : x, y \in \mathbb{R}\right\}.$$

It is clear that $\mathscr{V}$ is a set of functions on $\mathbb{R}$ and $\mathscr{V}_2$ is a set of functions on $\mathbb{R} \times \mathbb{R}$. Define an envelop function for $\mathscr{V}$ as follows. Since $\nu_F^n \to \nu_F$, by Skorohod representation theorem, there exists random variables X^n (with law ν_F^n) and X (with law ν_F), such that $X^n \to X$ almost surely as $r \to \infty$. Thus there exists a random variable X^* such that almost surely,

$$X^* = \sup_r X^n.$$

Let ν_F^* be the law of X^*. Since $L_2(\nu_F^*)$ (the space of square integrable functions with respect to the measure ν_F^*) contains continuous unbounded functions, there exists a continuous unbounded function $f_{\nu_F} : \mathbb{R}_+ \to \mathbb{R}$ that is increasing, satisfies $f_{\nu_F} \geq 1$ and $\langle f_{\nu_F}^2, \nu_F \rangle < \infty$. Similarly, based on the weak convergence $\nu_G^n \to \nu_G$, we can construct a function f_{ν_G} that is increasing, satisfies $f_{\nu_G} \geq 1$ and $\langle f_{\nu_G}^2, \nu_G \rangle < \infty$. Now, define function $\bar{f} : \mathbb{R}_+ \to \mathbb{R}$ by $\bar{f}(x) = \min\left(f_{\nu_F}(x), f_{\nu_G}(x)\right)$ and function $\bar{f}_2 : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ by $\bar{f}_2(x, y) = \min\left(f_{\nu_F}(x), f_{\nu_G}(y)\right)$ for all $x, y \in \mathbb{R}_+$. Note that we have to following properties,

$$\bar{f} \text{ is increasing and unbounded,} \tag{10.12}$$
$$f \leq \bar{f} \text{ for all } f \in \mathscr{V}, \tag{10.13}$$
$$f \leq \bar{f}_2 \text{ for all } f \in \mathscr{V}_2. \tag{10.14}$$

So we call $\bar{f}$ and $\bar{f}_2$ the envelop function for $\mathscr{V}$ and $\mathscr{V}_2$ respectively. Finally, let $\bar{\mathscr{V}} = \{\bar{f}\} \cup \mathscr{V}$ and $\bar{\mathscr{V}_2} = \{\bar{f}_2\} \cup \mathscr{V}_2$.

Lemma 10.1. *Assume that*

$$\nu_F^n \to \nu_F, \quad \nu_G^n \to \nu_G \text{ as } n \to \infty.$$

Fix constants $M, L > 0$. For all $\epsilon, \eta > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}} \left| \langle f, \bar{\mathcal{L}}_F^n(m,l)\rangle - l\langle f, \nu_F^n\rangle \right| > \epsilon \right) < \eta,$$

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}} \left| \langle f, \bar{\mathcal{L}}_G^n(m,l)\rangle - l\langle f, \nu_G^n\rangle \right| > \epsilon \right) < \eta,$$

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}_2} \left| \langle f, \bar{\mathcal{L}}_{F,G}^n(m,l)\rangle - l\langle f, (\nu_F^n, \nu_G^n)\rangle \right| > \epsilon \right) < \eta.$$

This kind of results have been widely used in the study of measure valued processes, see [24, 26, 63]. The proof of the first two inequalities in the above lemma follows exactly the same way as the one for Lemma $B.1$ in [63], and the proof of the third inequality in the above lemma follows exactly the same as the one for Lemma 5.1 in [26]. We omit the proof for brevity. By the same reasoning as for Lemma 4.1, there exists a function $\epsilon_{\mathrm{GC}}(\cdot)$, which vanishes at infinity such that the ϵ and η in the above lemma can be replaced by the function $\epsilon_{\mathrm{GC}}(n)$ for each index n. Based on this, we construct the following event,

$$\begin{aligned}\Omega_{\mathrm{GC}}^n(M,L) = &\left\{ \max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}} \left| \langle f, \bar{\mathcal{L}}_F^n(m,l)\rangle - l\langle f, \nu_F^n\rangle \right| \le \epsilon_{\mathrm{GC}}(n) \right\} \\ &\cap \left\{ \max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}} \left| \langle f, \bar{\mathcal{L}}_G^n(m,l)\rangle - l\langle f, \nu_G^n\rangle \right| \le \epsilon_{\mathrm{GC}}(n) \right\} \\ &\cap \left\{ \max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}_2} \left| \langle f, \bar{\mathcal{L}}_{F,G}^n(m,l)\rangle - l\langle f, (\nu_F^n, \nu_G^n)\rangle \right| \le \epsilon_{\mathrm{GC}}(n) \right\}. \end{aligned} \tag{10.15}$$

It is clear that for any fixed $M, L > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\Omega_{\mathrm{GC}}^n(M,L)\right) = 1. \tag{10.16}$$

Intuitively, on the event $\Omega_{\mathrm{GC}}^n(M,L)$ (whose probability goes to 1 as $n \to \infty$ for any fixed constants M, L), the measures $\bar{\mathcal{L}}_F^n(m,l)$, $\bar{\mathcal{L}}_G^n(m,l)$ and $\bar{\mathcal{L}}_{F,G}^n(m,l)$ are very "close" to $l\nu_F^n$, $l\nu_G^n$ and $l(\nu_F^n, \nu_G^n)$, respectively.

10.1.2 Compact Containment

A set $\mathbf{K} \subset \mathbf{M}$ is relatively compact if $\sup_{\xi \in \mathbf{K}} \xi(\mathbb{R}) < \infty$, and there exists a sequence of nested compact sets $A_j \subset \mathbb{R}$ such that $\cup A_j = \mathbb{R}$ and

$$\lim_{j \to \infty} \sup_{\xi \in \mathbf{K}} \xi(A_j^c) = 0,$$

where A_j^c denotes the complement of A_j; see [33], Theorem A7.5. The first major step to prove Theorem 10.2 is to establish the following *compact containment* property.

Lemma 10.2. *Assume* (4.8)–(4.15). *Fix* $T > 0$. *For each* $\eta > 0$ *there exists a compact set* $\mathbf{K} \subset \mathbf{M}$ *such that*

$$\liminf_{n \to \infty} \mathbb{P}\left((\bar{\mathcal{R}}^n(t), \bar{\mathcal{Z}}^n(t)) \in \mathbf{K} \times \mathbf{K} \text{ for all } t \in [0, T]\right) \geq 1 - \eta.$$

To prove this result, we first need to establish some bound estimations. For the convenience of notation, denote $\bar{E}^n(s,t) = \bar{E}^n(t) - \bar{E}^n(s)$ for any $0 \leq s \leq t$. Fix $T > 0$. It follows immediately from condition (4.8) that for each $\epsilon > 0$ there exists an n_0 such that when $n > n_0$,

$$\mathbb{P}\left(\sup_{0 \leq s < t \leq T} |\bar{E}^n(s,t) - \lambda(t-s)| < \epsilon\right) \geq 1 - \epsilon. \tag{10.17}$$

To facilitate some arguments later on, we derive the following result from the above inequality.

Lemma 10.3. *Fix* $T > 0$. *There exists a function* $\epsilon_E(\cdot)$, *with* $\lim_{n \to \infty} \epsilon_E(n) = 0$ *such that*

$$\mathbb{P}\left(\sup_{0 \leq s < t \leq T} |\bar{E}^n(s,t) - \lambda(t-s)| < \epsilon_E(n)\right) \geq 1 - \epsilon_E(n),$$

for each $n \geq 0$.

The derivation of the above lemma from (10.17) follows the same as the proof of Lemma 4.1. We omit the proof for brevity. Based on the above lemma, we construct the following event,

$$\Omega_E^n = \{\sup_{t \in [0,T]} |\bar{E}^n(s,t) - \lambda(t-s)| < \epsilon_E(n)\}. \tag{10.18}$$

We have that on this event, the arrival process is regular, i.e. $\bar{E}^n(s,t)$ is "close" to $\lambda(t-s)$. And this event has "large" probability, i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(\Omega_E^n\right) = 1. \tag{10.19}$$

Proof of Lemma 10.2. By the convergence of the initial condition (4.13), for any $\epsilon > 0$, there exists a relatively compact set $\mathbf{K}_0 \subset \mathbf{M}$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\bar{\mathcal{R}}^n(0) \in \mathbf{K}_0 \text{ and } \bar{\mathcal{Z}}^n(0) \in \mathbf{K}_0\right) > 1-\epsilon. \tag{10.20}$$

Denote the event in the above probability by Ω_0^n. On this event, by the definition of relatively compact set in the space $\mathbf{M}$, there exists a function $\kappa_0(\cdot)$ with $\lim_{x\to\infty}\kappa_0(x) = 0$ such that

$$\bar{\mathcal{R}}^n(0)(C_x) \le \kappa_0(x), \quad \bar{\mathcal{Z}}^n(0)(C_x) \le \kappa_0(x), \tag{10.21}$$

and

$$\bar{\mathcal{R}}^n(0)(C_x^-) \le \kappa_0(x), \tag{10.22}$$

for all $x \ge 0$, where $C_x^- = (-\infty, -x)$ for any $y \in \mathbb{R}$. (Remember that $\bar{\mathcal{Z}}^n(0)$ is a measure on $\mathbb{R}_+$, so we do not need to consider its measure of C_x^-.) It is clear that on the event $\Omega_E^n \cap \Omega_0^n$, for any $t \le T$ and all large n,

$$\begin{aligned}
\bar{\mathcal{R}}^n(t)(\mathbb{R}) &\le \sup_n \bar{\mathcal{R}}^n(0)(\mathbb{R}) + 2\lambda T,\\
\bar{\mathcal{Z}}^n(t)(\mathbb{R}_+) &\le 1,
\end{aligned}$$

where the last inequality is due to the fact that $Z^n(\cdot) \le n$. Again, by the definition of relative compact set in $\mathbf{M}$, we have that $\sup_n \bar{\mathcal{R}}^n(0)(\mathbb{R}) = M_0 < \infty$. It follows from the dynamic equation (10.7) and (10.8) that for all $x > 0$,

$$\begin{aligned}
\bar{\mathcal{R}}^n(t)(C_x) &\le \bar{\mathcal{R}}^n(0)(C_x) + \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{u_i^n}(C_x),\\
\bar{\mathcal{Z}}^n(t)(C_x) &\le \bar{\mathcal{Z}}^n(0)(C_x) + \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{v_i^n}(C_x).
\end{aligned}$$

Denote $\bar{\mathcal{L}}_1^n(t) = \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{u_i^n}$ and $\bar{\mathcal{L}}_2^n(t) = \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{v_i^n}$. Let us first study these two terms. Recall the definition of the event $\Omega_{\mathrm{GC}}^n(M, L)$ and the envelope function $\bar{f}$ (which increases to infinity) in (10.15). For the application here, it is enough to set $M = 1$ and $L = 2\lambda T$. On the event $\Omega_E^n \cap \Omega_{\mathrm{GC}}^n(M, L)$, we have

$$\langle \bar{f}, \bar{\mathcal{L}}_1^n(t)\rangle \le \langle \bar{f}, \frac{1}{n}\sum_{i=1}^{2\lambda T n} \delta_{u_i^n}\rangle \le 2\lambda T \langle \bar{f}, \nu_F\rangle + 1,$$

for all large enough n. Similarly, on the same event we have that

$$\langle \bar{f}, \bar{\mathcal{L}}_2^n(t)\rangle \le \langle \bar{f}, \frac{1}{n}\sum_{i=1}^{2\lambda T n} \delta_{v_i^n}\rangle \le 2\lambda T \langle \bar{f}, \nu_G\rangle + 1,$$

for all large enough n. Denote $M_B = 2\lambda T \max(\langle \bar{f}, \nu_F\rangle, \langle \bar{f}, \nu_G\rangle) + 1$. By Markov's inequality, for all $x > 0$ (again, on the same event and for all large n)

$$\bar{\mathcal{L}}_1^n(t)(C_x) < M_b/\bar{f}(x), \quad \bar{\mathcal{L}}_2^n(t)(C_x) < M_b/\bar{f}(x).$$

Unlike the measure $\mathcal{Z}(t) \in \mathbf{M}_+$, the measure $\mathcal{R}(t) \in \mathbf{M}$. So we need to consider all the test set $C_x^- = (-\infty, -x)$ for $x \ge 0$. The following inequality again follows from (10.7),

$$\bar{\mathcal{R}}^n(t)(C_x^-) \le \bar{\mathcal{R}}^n(0)(C_x^- + t) + \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{u_i^n}(C_x^- + t).$$

Note that if we take $x > T$, then $\delta_{u_i^n}(C_x^- + t) = 0$. So we have that

$$\bar{\mathcal{R}}^n(t)(C_x^-) \le \bar{\mathcal{R}}^n(0)(C_x^- + T) = \bar{\mathcal{R}}^n(0)(C_{x-T}^-), \quad \text{for all } t \le T. \tag{10.23}$$

Now, define the set $\mathbf{K} \subset \mathbf{M}$ by

$$\begin{aligned}\mathbf{K} = \Big\{\xi \in \mathbf{M} :& \xi(\mathbb{R}) < 1 + M_0 + 2\lambda T,\\ & \xi(C_x) < \kappa_0(x) + M_b/\bar{f}(x) \text{ for all } x > 0,\\ & \xi(C_x^-) \le \kappa_0(x - T) \text{ for all } x \ge T\Big\}.\end{aligned}$$

It is clear that $\mathbf{K}$ is relatively compact and on the event $\Omega_E^n \cap \Omega_{\mathrm{GC}}^n(M, L) \cap \Omega_0^n$,

$$(\bar{\mathcal{R}}^n(t), \bar{\mathcal{Z}}^n(t)) \in \mathbf{K} \times \mathbf{K} \text{ for all } t \in [0, T].$$

The result of this lemma then follows immediately from (10.19), (10.20) and (10.16).

□

10.1.3 Oscillation Bound

The second major step to prove precompactness is to obtain the oscillation bound in Lemma 10.5 below. The oscillation of a *càdlàg* function $\zeta(\cdot)$ (taking values in a metric space $(\mathbf{E}, \pi)$) on a fixed interval $[0, T]$ is defined as

$$\mathbf{w}_L(\zeta(\cdot), \delta)T = \sup_{s,t\in[0,T],|s-t|<\delta} \pi[\zeta(s), \zeta(t)].$$

If the metric space is $\mathbb{R}$, we just use the Euclidean metric; if the space is $\mathbf{M}$ or $\mathbf{M}_+$, we use the Prohorov metric $\mathbf{d}$ defined in Section 1.4. For the measure-valued processes in our model, oscillations mainly result from sudden departures of a large number of customers. To control the departure process, we show that $\bar{\mathcal{Z}}^n(\cdot)$ and $\bar{\mathcal{R}}^n(\cdot)$ assign arbitrarily small mass to small intervals.

Lemma 10.4. *Assume* (9.10), (10.2)–(10.5). *Fix* $T > 0$. *For each* $\epsilon, \eta > 0$ *there exists a* $\kappa > 0$ *(depending on* ϵ *and* η*) such that*

$$\liminf_{n\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^n(t)([x, x+\kappa]) \le \epsilon\right) \ge 1 - \eta. \tag{10.24}$$

Proof. First, We have that for any $\epsilon, \eta > 0$, there exists a κ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^n(0)([x, x+\kappa]) \le \epsilon/2\right) \ge 1 - \eta. \tag{10.25}$$

This inequality is derived from the initial condition. The derivation is exactly the same as in the proof of (5.14) in [63], so we omit it here for brevity.

Now we need to extend this result to the interval $[0, T]$. Denote the event in (10.25) by Ω_0^n, and the event in Lemma 10.2 by $\Omega_C^n(\mathbf{K})$. Fix $M = 1$ and $L = 2\lambda T$, Let

$$\Omega_1^n(M, L) = \Omega_0^n \cap \Omega_C^n(\mathbf{K}) \cap \Omega_E^n \cap \Omega_{\mathrm{GC}}^n(M, L). \tag{10.26}$$

By (10.25), Lemma 10.2, (10.19) and (10.16), for any fixed $M, L > 0$,

$$\liminf_{n\to\infty} \mathbb{P}\left(\Omega_1^n(M, L)\right) \ge 1 - \eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega_1^n(M, L)$.

It follows from the fluid scaled stochastic dynamic equation (10.8) that

$$\bar{\mathcal{Z}}^n(t)([x, x+\kappa]) \leq \bar{\mathcal{Z}}^n(0)([x, x+\kappa] + t) + \frac{1}{n} \sum_{i=B(0)+1}^{B(t)} \delta_{v_i^n}([x, x+\kappa] + t - \tau_i^n),$$

for each $x, \kappa \in \mathbb{R}_+$. By (10.25), the first term on the right hand side of the above equation is always upper bounded by $\epsilon/2$. Let S denote the second term on the right hand side of the preceding equation. Now it only remains to show that $S < \epsilon/2$.

Let $0 = t_0 < t_1 < \cdots < t_J = t$ be a partition of the interval $[0, t]$ such that $|t_{j+1} - t_j| < \delta$ for all $j = 0, \cdots, J-1$, where δ and N are to be chosen below. Write S as the summation

$$S = \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=B(t_j)+1}^{B(t_{j+1})} \delta_{v_i^n}([x, x+\kappa] + t - \tau_i^n).$$

Recall that τ_i^n is the time that the ith job starts service, so on each sub-interval $[t_j, t_{j+1}]$ those i's to be summed must satisfy $t_j \leq \tau_i^n \leq t_{j+1}$. This implies that

$$t - t_{j+1} \leq t - \tau_i \leq t - t_j.$$

Then

$$S \leq \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=B(t_j)+1}^{B(t_{j+1})} \delta_{v_i^n}([x + t - t_{j+1}, x + t - t_j + \kappa]).$$

By (10.6), we have for all $j = 0, \cdots, J-1$

$$-\bar{R}^n(0) \leq \bar{B}^n(t_j) \leq \bar{E}^n(T),$$

$$0 \leq \bar{B}^n(t_{j+1}) - \bar{B}^n(t_j) \leq \bar{E}^n(T) + \bar{R}^n(0).$$

By Lemmas 10.2 and 4.1, $\bar{R}^n(0) < M_0$ and $\bar{E}^n(T) \leq 2\lambda T$ on $\Omega_C^n(\mathbf{K}) \cap \Omega_E^n$ for some constant M_0. Take $M = \max(M_0, 2\lambda T)$ and $L = M_0 + 2\lambda T$, it follows from the

Glevenko-Cantelli estimate (10.15) that

$$\frac{1}{n}\sum_{i=B^n(t_j)+1}^{B^n(t_{j+1})}\delta_{v_i^n}([x+t-t_{j+1},x+t-t_j+\kappa])$$
$$\leq \Big(\bar{B}^n(t_{j+1})-\bar{B}^n(t_j)\Big)\nu^n([x+t-t_{j+1},x+t-t_j+\kappa])+\frac{\epsilon}{4J},$$

for each $j<J$. By condition (10.3), for any $\epsilon_2>0$,

$$\mathbf{d}[\nu_G^n,\nu_G]<\epsilon_2,$$

for all large n. By the definition of Prohorov metric, we have

$$\nu_G^n([x+t-t_{j+1},x+t-t_j+\kappa])\leq \nu_G([x+t-t_{j+1}-\epsilon_2,x+t-t_j+\kappa+\epsilon_2]),$$

for all large n. Since $[x+t-t_{j+1}-\epsilon_2,x+t-t_j+\kappa+\epsilon_2]$ is a close interval with length less than $\kappa+\delta+2\epsilon_2$, by condition (9.10), we can choose κ,δ,ϵ_2 small enough such that

$$\nu([x+t-t_{j+1}-\epsilon_2,x+t-t_j+\kappa+\epsilon_2])\leq\frac{\epsilon}{4M}.$$

Thus, we conclude that

$$S\leq\frac{\epsilon}{4J}[\bar{B}^n(T)-\bar{B}^n(0)]+\frac{\epsilon}{4}\leq\epsilon/2.$$

This completes the proof. □

Lemma 10.5. *Assume* (9.10), (10.2)–(10.5). *Fix* $T>0$. *For each* $\epsilon,\eta>0$ *there exists a* $\delta>0$ *(depending on* ϵ *and* η*) such that*

$$\liminf_{n\to\infty}\mathbb{P}\left(\mathbf{w}_L((\bar{\mathcal{R}}^n,\bar{\mathcal{Z}}^n)(\cdot),\delta)T\leq 3\epsilon\right)\geq 1-\eta. \tag{10.27}$$

Proof. Define

$$\Omega^n_{\mathrm{Reg}}(\epsilon,\kappa)=\Big\{\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\bar{\mathcal{Z}}^n(t)([x,x+\kappa])\leq\epsilon\Big\}.$$

By (10.19) and Lemma 10.4, for each $\epsilon,\eta>0$ there exists a $\kappa>0$ such that

$$\liminf_{n\to\infty}\mathbb{P}\left(\Omega_E^n\cap\Omega^n_{\mathrm{Reg}}(\epsilon,\kappa)\right)>1-\eta. \tag{10.28}$$

On the event $\Omega_E^n \cap \Omega_{\text{Reg}}^n(\epsilon, \kappa)$, we have some control over the dynamics of the system. First, note that the number of customers (in the virtual buffer, including those who have abandoned but ought to get service if they did not) that enter the server during time interval $(s, t]$ can be upper bounded by

$$\bar{B}^n(s,t) \leq \bar{E}^n(s,t) + \bar{\mathcal{Z}}^n(s)([0, t-s]).$$

When $t - s \leq \min(\frac{\epsilon}{2\lambda}, \kappa)$, by the definition of Ω_E^n and $\Omega_{\text{Reg}}^n(\epsilon, \kappa)$, we have

$$\bar{E}^n(s,t) \leq \epsilon \tag{10.29}$$

$$\bar{B}^n(s,t) \leq 2\epsilon. \tag{10.30}$$

Second, by the dynamic equation (10.7), for any $s < t$ and any set $C \in \mathscr{B}(\mathbb{R})$,

$$\begin{aligned}\bar{\mathcal{R}}^n(t)(C) - \bar{\mathcal{R}}^n(s)(C^{3\epsilon})) \leq \bar{B}^n(s,t) + \bar{E}^n(s,t) \\ + \frac{1}{n} \sum_{1+B^n(t)}^{E^n(s)} [\delta_{u_i^n}(C + t - a_i^n) - \delta_{u_i^n}(C^{3\epsilon} + s - a_i^n)],\end{aligned}$$

where C^a is the a-enlargement of the set C as defined in Section 1.4. Note that when $t - s \leq 3\epsilon$, $C + t - a_i^n \subseteq C^{3\epsilon} + s - a_i^n$ for all $i \in \mathbb{Z}$, which implies that the second term in the above inequality is less than zero. By (10.29) and (10.30),

$$\bar{\mathcal{R}}^n(t)(C) - \bar{\mathcal{R}}^n(s)(C^{3\epsilon})) \leq 3\epsilon.$$

By Property (ii) on page 72 in [5], we have

$$\mathbf{d}[\bar{\mathcal{R}}^n(t), \bar{\mathcal{R}}^n(s)] \leq 3\epsilon. \tag{10.31}$$

Finally, by the dynamic equation (10.8),

$$\bar{\mathcal{Z}}^n(t)(C) \leq \bar{\mathcal{Z}}^n(s)(C + t - s)) + \bar{B}^n(s,t).$$

Note that when $t - s \leq 2\epsilon$, $C + t - s \subseteq C^{2\epsilon}$, where C^a is the a-enlargement of the set C as defined in Section 1.4. By (10.30), we have

$$\bar{\mathcal{Z}}^n(t)(C) \leq \bar{\mathcal{Z}}^n(s)(C^{2\epsilon}) + 2\epsilon.$$

By Property (ii) on page 72 in [5], we have

$$\mathbf{d}[\bar{\mathcal{Z}}^n(s), \bar{\mathcal{Z}}^n(t)] \leq 2\epsilon. \tag{10.32}$$

The result of this lemma follows immediately from (10.28), (10.31) and (10.32). □

10.2 Convergence to the Fluid Model Solution

We have established the precompactness in Theorem 10.2. So every subsequence of the fluid scaled processes has a further subsequence which converges to some limit. For simplicity of notations, we index the convergent subsequence again by n. So we have that

$$(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)) \Rightarrow (\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \quad \text{as } n \to \infty. \tag{10.33}$$

By the oscillation bound in Lemma 10.5, the limit $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is almost surely continuous. We have the following result that further characterizes the above limit.

Lemma 10.6. *Assume* (9.10)–(9.13) *and* (10.2)–(10.5). *The limit* $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ *in* (10.33) *is almost surely the solution to the fluid model* (λ, H) *with initial condition* $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

The rest of this section is devoted to characterizing the limits. To better structure the proof, we first provide some preliminary estimates based on the dynamic equations (10.7) and (10.8).

Lemma 10.7. *Let* $\{t_j\}_{j=0}^J$ *be a partition of the interval* $[s,t]$ *such that* $s = t_0 < t_1 < \ldots < t_J = t$. *We have for any* $x \in \mathbb{R}$,

$$\bar{\mathcal{R}}^n(t)(C_x) \leq \sum_{i=0}^{J-1} \frac{1}{n} \sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})} \delta_{u_i^n}(C_x + t - t_j) + |\bar{E}^n(s) - \bar{B}^n(t)|, \tag{10.34}$$

$$\bar{\mathcal{R}}^n(t)(C_x) \geq \sum_{i=0}^{J-1} \frac{1}{n} \sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})} \delta_{u_i^n}(C_x + t - t_{j+1}) - |\bar{E}^n(s) - \bar{B}^n(t)|. \tag{10.35}$$

If in addition that $\sup_{\tau\in[s,t]}|\bar{E}^n(\tau)-\lambda\tau|<\epsilon$, *then for any* $x>0$,

$$\bar{\mathcal{Z}}^n(t)(C_x)\le\bar{\mathcal{Z}}^n(s)(C_x+t-s)+\sum_{j=0}^{J-1}\frac{1}{n}\sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})}\delta_{u_i^n}(C_0+\frac{\bar{R}^n_{L,j}-2\epsilon}{\lambda})\delta_{v_i^n}(C_x+t-t_j), \tag{10.36}$$

$$\bar{\mathcal{Z}}^n(t)(C_x)\ge\bar{\mathcal{Z}}^n(s)(C_x+t-s)+\sum_{j=0}^{J-1}\frac{1}{n}\sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})}\delta_{u_i^n}(C_0+\frac{\bar{R}^n_{U,j}+2\epsilon}{\lambda})\delta_{v_i^n}(C_x+t-t_{j+1}), \tag{10.37}$$

where $\bar{R}^n_{L,j}=\inf_{t\in[t_j,t_{j+1}]}\bar{R}^n(t)$ *and* $\bar{R}^n_{U,j}=\sup_{t\in[t_j,t_{j+1}]}\bar{R}^n(t)$.

Proof. Note that $0\le\delta_{u_i^n}(C)\le 1$ for any Borel set C and any random variable u_i^n. So by the dynamic equation (10.7), we have

$$\Big|\bar{\mathcal{R}}^n(t)(C)-\frac{1}{n}\sum_{i=E^n(s)+1}^{E^n(t)}\delta_{u_i^n}(C+t-a_i^n)\Big|\le|\bar{E}^n(s)-\bar{B}^n(t)|.$$

For those i's such that $E^n(t_j)<i\le E^n(t_{j+1})$, we have that

$$t_j<a_i^n\le t_{j+1}. \tag{10.38}$$

This implies that $C_x+t-a_i\subseteq C_x+t-t_j$. So we have

$$\sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})}\delta_{u_i^n}(C_x+t-a_i)\le\sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})}\delta_{u_i^n}(C_x+t-t_j).$$

This establishes (10.34). Also, (10.38) implies $C_x+t-t_{j+1}\subseteq C_x+t-a_i$. So (10.35) follows in the same way.

For those i's such that $B^n(t_j)<i\le B^n(t_{j+1})$, we have that

$$t_j<\tau_j^n\le t_{j+1}.$$

Note that $\bar{R}^n(\tau_i^n)=\bar{E}^n(\tau_i^n)-\bar{E}^n(a_i^n)$ for each i. So, by the closeness between $\bar{E}^n(\cdot)$ and $\lambda\cdot$, we have

$$\begin{aligned}&|\bar{R}^n(\tau_i^n)-\lambda(\tau_i^n-a_i^n)|\\ &\le|\bar{R}^n(\tau_i^n)-\bar{E}^n(\tau_i^n)+\bar{E}^n(a_i^n)|+|\bar{E}^n(\tau_i^n)-\bar{E}^n(a_i^n)-\lambda(\tau_i^n-a_i^n)|\\ &\le 2\epsilon.\end{aligned}$$

So

$$\bar{R}^n_{L,j} - 2\epsilon \le \lambda(\tau^n_i - a^n_i) \le \bar{R}^n_{U,j} + 2\epsilon,$$

for all i's such that $B^n(t_j) < i \le B^n(t_{j+1})$. Thus,

$$\sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})} \delta_{u^n_i}(C_0+\tau^n_i-a^n_i)\delta_{v^n_i}(C_x+t-\tau^n_j) \le \sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})} \delta_{u^n_i}(C_0+\frac{\bar{R}^n_{L,j}-2\epsilon}{\lambda})\delta_{v^n_i}(C_x+t-t_j).$$

This implies (10.36). And (10.37) can be proved in the same way. □

Recall the notations $\bar{\mathcal{L}}^n(m,l), \bar{\mathcal{L}}^n_p(m,l)$ and $\bar{\mathcal{L}}^n_S(m,l)$ are defined in (10.9)–(10.11) in the appendix. Using these notations, Lemma 10.7 can be written as the following:

Lemma 10.8. *Let $\{t_j\}_{j=0}^J$ be a partition of the interval $[s,t]$ such that $s = t_0 < t_1 < \ldots < t_J = t$. We have for any $x \in \mathbb{R}$,*

$$\bar{\mathcal{R}}^n(t)(C_x) \le \sum_{i=0}^{J-1} \langle 1_{(C_x+t-t_j)}, \bar{\mathcal{L}}^n_p(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\rangle + |\bar{E}^n(s) - \bar{B}^n(t)|, \tag{10.39}$$

$$\bar{\mathcal{R}}^n(t)(C_x) \ge \sum_{i=0}^{J-1} \langle 1_{(C_x+t-t_{j+1})}, \bar{\mathcal{L}}^n_p(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\rangle - |\bar{E}^n(s) - \bar{B}^n(t)|. \tag{10.40}$$

If in addition that $\sup_{\tau\in[s,t]} |\bar{E}^n(\tau) - \lambda\tau| < \epsilon$, then for any $x > 0$,

$$\begin{aligned}\bar{\mathcal{Z}}^n(t)(C_x) \le\ & \bar{\mathcal{Z}}^n(s)(C_x + t - s) \\ & + \sum_{j=0}^{J-1} \langle 1_{(C_0+\frac{\bar{R}^n_{L,j}-2\epsilon}{\lambda})\times(C_x+t-t_j)}, \bar{\mathcal{L}}^n(B^n(t_j), \bar{B}^n(t_j,t_{j+1}))\rangle,\end{aligned} \tag{10.41}$$

$$\begin{aligned}\bar{\mathcal{Z}}^n(t)(C_x) \ge\ & \bar{\mathcal{Z}}^n(s)(C_x + t - s) \\ & + \sum_{j=0}^{J-1} \langle 1_{(C_0+\frac{\bar{R}^n_{U,j}+2\epsilon}{\lambda})\times(C_x+t-t_{j+1})}, \bar{\mathcal{L}}^n(B^n(t_j), \bar{B}^n(t_j,t_{j+1}))\rangle.\end{aligned} \tag{10.42}$$

Fix a constant $T > 0$ and let $M = 1$ and $L = 2\lambda T$. Denote the random variable

$$\bar{V}^n_{M,L} = \max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{x,y\in\mathbb{R}} \left\{ \begin{array}{l} \left|\bar{\mathcal{L}}^n(m,l)(C_x\times C_y) - l\nu^n_F(C_x)\nu^n_G(C_y)\right| \\ +\left|\bar{\mathcal{L}}^n_F(m,l)(C_x) - l\nu^n_F(C_x)\right| \\ +\left|\bar{\mathcal{L}}^n_G(m,l)(C_x) - l\nu^n_G(C_x)\right| \end{array} \right\}. \tag{10.43}$$

By Lemma D.2, for any fixed constants $M, L > 0$,

$$\bar{V}^n_{M,L} \Rightarrow 0 \quad \text{as } n \to \infty.$$

By the assumption (10.2), we have

$$\bar{E}^n(\cdot) \Rightarrow \lambda \cdot \quad \text{as } n \to \infty.$$

Since both the above two limits are deterministic, those convergences are joint with the convergence of $(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot))$. Now, for each $n \geq 1$, we can view $(\bar{E}^n(\cdot), \bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot), V_{M,L})$ as a random variable in the space $\mathbf{E}_1$, which is the product space of three $\mathbf{D}([0,\infty), \mathbb{R})$ spaces and the space $\mathbb{R}$. And $(\bar{\mathcal{L}}^n(m,\cdot), \bar{\mathcal{L}}^n_F(m,\cdot), \bar{\mathcal{L}}^n_G(m,\cdot) : m \in \mathbb{Z})$ in the product space $\mathbf{E}_2$ of countable many $\mathbf{D}([0,\infty), \mathbf{M})$ spaces. It is clear that both $\mathbf{E}_1$ and $\mathbf{E}_2$ are complete and separable metric spaces. Using the extension of Skorohod representation Theorem, Lemma F.1, we assume without loss of generality that $\bar{E}^n(\cdot), \bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot), \bar{V}^n_{M,L}, \bar{\mathcal{L}}^n(m,\cdot), \bar{\mathcal{L}}^n_F(m,\cdot), \bar{\mathcal{L}}^n_G(m,\cdot), m \in \mathbb{Z}$, and $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ are defined on a common probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that, almost surely,

$$\Big((\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)), \bar{V}^n_{M,L}, \bar{E}^n(\cdot)\Big) \to \Big((\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)), 0, \lambda \cdot \Big) \quad \text{as } n \to \infty, \tag{10.44}$$

and inequalities (10.39)–(10.42) and equation (10.43) also hold almost surely. Note that the convergence of each function component in the above is in the Skorohod J_1 topology. Since the limit is continuous, the convergence is equivalent to the convergence in the uniform norm on compact intervals. Thus as $n \to \infty$,

$$\sup_{t\in[0,T]} \mathbf{d}[\bar{\mathcal{R}}^n(t), \tilde{\mathcal{R}}(t)] \to 0, \tag{10.45}$$

$$\sup_{t\in[0,T]} \mathbf{d}[\bar{\mathcal{Z}}^n(t), \tilde{\mathcal{Z}}(t)] \to 0, \tag{10.46}$$

$$\sup_{t\in[0,T]} \left|\bar{E}^n(t) - \lambda t\right| \to 0. \tag{10.47}$$

Same as on the original probability space, let

$$\bar{R}^n(\cdot) = \langle 1, \bar{\mathcal{R}}^n(\cdot)\rangle, \quad \bar{Q}^n(\cdot) = \langle 1_{(0,\infty)}, \bar{\mathcal{R}}^n(\cdot)\rangle,$$
$$\bar{Z}^n(\cdot) = \langle 1, \bar{\mathcal{Z}}^n(\cdot)\rangle, \quad \bar{X}^n(\cdot) = \bar{Q}^n(\cdot) + \bar{Z}^n(\cdot),$$

and

$$\bar{B}^n(\cdot) = \bar{E}^n(\cdot) - \bar{R}^n(\cdot).$$

According to (10.45) and (10.47), we have

$$\sup_{t\in[0,T]} \left|\bar{B}^n(t) - \tilde{B}(t)\right| \to 0. \tag{10.48}$$

For each n, let $\tilde{\Omega}_{n,2}$ be an event of probability one on which the stochastic dynamic equations (10.7) and (10.8) and the policy constraints (8.7) and (8.8) hold. Define $\tilde{\Omega}_0 = \tilde{\Omega}_1 \cap (\cap_{n=0}^{\infty} \tilde{\Omega}_{n,2}^n)$, where $\tilde{\Omega}_1$ is the event of probability one on which (10.44) holds. Then $\tilde{\Omega}_0$ also has probability one. Based on Lemma 10.7 and the above argument using Skorohod Representation theorem, we can now prove Lemma 10.6.

Proof of Lemma 10.6. For any $t \geq 0$, fix a constant $T > t$. Let us now study $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ on the time interval $[0, T]$. It is enough to show that on the event $\tilde{\Omega}_0$, $(\tilde{\mathcal{R}}(t), \tilde{\mathcal{Z}}(t))$ satisfies the fluid model equation (9.1)–(9.2) and the constraints (9.3)–(9.4). Assume for the remainder of this proof that all random objects are evaluated at a sample path in the event $\tilde{\Omega}_0$.

We first verify (9.1). For any $\epsilon > 0$, consider the difference

$$\begin{aligned}
&\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)d\lambda s \\
&= \tilde{\mathcal{R}}(t)(C_x) - \bar{\mathcal{R}}^n(t)(C_x^\epsilon) + \bar{\mathcal{R}}^n(t)(C_x^\epsilon) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)d\lambda s,
\end{aligned}$$

where C_x^ϵ is the ϵ-enlargement of the set C_x as defined in Chapter 1.4, which is essentially $C_{x-\epsilon}$. Let $t_0 = t - \tilde{R}(t)/\lambda$. According to (10.39), we have that

$$\begin{aligned}
&\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)d\lambda s \\
&\leq \tilde{\mathcal{R}}(t)(C_x) - \bar{\mathcal{R}}^n(t)(C_x^\epsilon) + |\bar{E}^n(t_0) - \bar{B}^n(t)| \\
&\sum_{i=0}^{J-1} \langle 1_{(C_x^\epsilon + t - t_j)}, \bar{\mathcal{L}}_p^n(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\rangle - \int_{t_0}^{t} F^c(x+t-s)d\lambda s,
\end{aligned} \tag{10.49}$$

where $\{t_j\}_{j=0}^J$ is a partition of the interval $[t_0, t]$ such that $t_0 < t_1 < \ldots < t_J = t$ and $\max_j(t_{j+1} - t_j) < \delta$ for some $\delta > 0$. By the definition of Prohorov metric and the convergence in (10.45), the first term on the right hand side of (10.49) is bounded by ϵ for all large n. By (10.45) and (10.47)

$$\begin{aligned}|\bar{B}^n(t) - \bar{E}^n(t_0)| &= |\bar{E}^n(t) - \bar{R}^n(t) - \bar{E}^n(t_0)| \\ &\leq |\bar{E}^n(t) - \lambda t| + |\bar{R}^n(t) - \tilde{R}(t)| + |\bar{E}^n(t_0) - \lambda t_0| < 3\epsilon,\end{aligned}$$

for all large n. So

$$\begin{aligned}&\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^t F^c(x+t-s)d\lambda s \\ &\leq 4\epsilon + \sum_{i=0}^{J-1}\langle 1_{(C_x^\epsilon + t - t_j)}, \bar{\mathcal{L}}_p^n(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\rangle - \int_{t_0}^t F^c(x+t-s)d\lambda s,\end{aligned} \tag{10.50}$$

for all large n. Similarly, according to (10.40), we have

$$\begin{aligned}&\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^t F^c(x+t-s)d\lambda s \\ &\geq -4\epsilon + \sum_{i=0}^{J-1}\langle 1_{(C_x^\epsilon + t - t_{j+1})}, \bar{\mathcal{L}}_p^n(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\rangle - \int_{t_0}^t F^c(x+t-s)d\lambda s,\end{aligned} \tag{10.51}$$

for all large n. Note that for each j, we have

$$\begin{aligned}&\langle 1_{(C_x + t - t_j)}, \bar{\mathcal{L}}_p^n(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\rangle \\ &\leq \langle 1_{(C_x + t - t_j)}, \bar{\mathcal{L}}_p^n(E^n(t_j), \lambda(t_{j+1} - t_j) + 2\epsilon\rangle \\ &\leq [\lambda(t_{j+1} - t_j) + 2\epsilon]\nu_F^n(C_x^\epsilon + t - t_j) + \epsilon \\ &\leq [\lambda(t_{j+1} - t_j) + 2\epsilon][\nu_F(C_x + t - t_j) + \epsilon] + \epsilon \\ &\leq \lambda(t_{j+1} - t_j)\nu_F(C_x + t - t_j) + (3 + \lambda\delta)\epsilon\end{aligned}$$

for all large n, where the first inequality is due to (10.47), the second one is due to (10.44) (the component of $\bar{V}_{M,L}^n$), the third one is due to (10.3), and the last one is due to algebra. Similarly, we can show that

$$\begin{aligned}&\langle 1_{(C_x + t - t_{j+1})}, \bar{\mathcal{L}}_p^n(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\rangle \\ &\geq \lambda(t_{j+1} - t_j)\nu_F(C_x + t - t_{j+1}) - (3 + \lambda\delta)\epsilon\end{aligned}$$

for all large n. Note that $\sum_{j=0}^{J-1} \lambda(t_{j+1}-t_j)F^c(x+t-t_j)$ and $\sum_{j=0}^{J-1} \lambda(t_{j+1}-t_j)F^c(x+t-t_{j+1})$ serve as the upper and lower Reimann sum of the integral $\int_{t_0}^{t} F^c(x+t-s)d\lambda s$, which converge to the integration as $n \to \infty$. So by (10.50) and (10.51), we have that for all large n,

$$\left|\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)d\lambda s\right| \leq (3+\lambda\delta)J\epsilon + 5\epsilon.$$

We conclude that $\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)d\lambda s = 0$ since ϵ in the above can be arbitrary. This verifies (9.1).

Next, we verify (9.2). For any $\epsilon > 0$, consider the difference

$$\begin{aligned}
&\left|\tilde{\mathcal{Z}}(t)(C_x) - \bar{\mathcal{Z}}_0(C_x+t) - \int_0^t F^c(\frac{\tilde{R}(s)}{\lambda})G^c(x+t-s)d[\lambda s - \tilde{R}(s)]\right| \\
&\leq |\tilde{\mathcal{Z}}(t)(C_x) - \bar{\mathcal{Z}}^n(t)(C_x^\epsilon)| + |\tilde{\mathcal{Z}}_0(C_x+t) - \bar{\mathcal{Z}}^n(0)(C_x^\epsilon+t)| \\
&\quad + \left|\bar{\mathcal{Z}}^n(t)(C_x^\epsilon) - \bar{\mathcal{Z}}^n(0)(C_x^\epsilon+t) - \int_0^t F^c(\frac{\tilde{R}(s)}{\lambda})G^c(x+t-s)d[\lambda s - \tilde{R}(s)]\right|,
\end{aligned} \tag{10.52}$$

where the above inequality is due to the fluid scaled stochastic dynamic equation (10.8). Again, by the definition of Prohorov metric and the convergence in (10.46), each of the first two terms on the right hand side in the above inequality is less than ϵ for all large n. Let $\{t_j\}_{j=0}^J$ be a partition of the interval $[0,t]$ such that $0 = t_0 < t_1 < \ldots < t_J = t$ and $\max_j(t_{j+1}-t_j) < \delta$ for some $\delta > 0$. Let

$$\tilde{R}_{U,j} = \sup_{t\in[t_j,t_{j+1}]} \tilde{R}(t), \quad \tilde{R}_{L,j} = \inf_{t\in[t_j,t_{j+1}]} \tilde{R}(t).$$

By (10.45), we have that

$$|\bar{R}^n_{U,j} - \tilde{R}_{U,j}| \leq \epsilon, \quad |\bar{R}^n_{L,j} - \tilde{R}_{L,j}| \leq \epsilon,$$

for all large n. So for each j, we have

$$\begin{aligned}
&\langle 1_{(C_0+\frac{\bar{R}^n_{L,j}-2\epsilon}{\lambda})\times(C^\epsilon_x+t-t_j)}, \bar{\mathcal{L}}^n(B^n(t_j), \bar{B}^n(t_j,t_{j+1}))\rangle \\
\leq\ &\langle 1_{(C_0+\frac{\tilde{R}_{L,j}-3\epsilon}{\lambda})\times(C^\epsilon_x+t-t_j)}, \bar{\mathcal{L}}^n(B^n(t_j), \tilde{B}(t_{j+1})-\tilde{B}(t_j)+2\epsilon)\rangle \\
\leq\ &[\tilde{B}(t_{j+1})-\tilde{B}(t_j)+2\epsilon]\nu^n_F(C_0+\frac{\tilde{R}_{L,j}-3\epsilon}{\lambda})\nu^n_G(C^\epsilon_x+t-t_j)+\epsilon \\
\leq\ &[\tilde{B}(t_{j+1})-\tilde{B}(t_j)+2\epsilon][\nu_F(C_0+\frac{\tilde{R}_{L,j}}{\lambda})+\frac{3\epsilon}{\lambda}][\nu_G(C_x+t-t_j)+\epsilon]+\epsilon
\end{aligned}$$

for all large n, where the first inequality is due to (10.48), the second one is due to (10.44) (the component of $\bar{V}^n_{M,L}$), the third one is due to (10.3). Let M_B be a finite upper bound of $\tilde{B}(t_J)-\tilde{B}(t_0)$, the above inequality can be further bounded by

$$[\tilde{B}(t_{j+1})-\tilde{B}(t_j)]\nu_F(C_0+\frac{\tilde{R}_{L,j}}{\lambda})\nu_G(C_x+t-t_j)+(\frac{3}{\lambda}+2)M_B\epsilon+3\epsilon.$$

Similarly, we can show that

$$\begin{aligned}
&\langle 1_{(C_0+\frac{\bar{R}^n_{U,j}+2\epsilon}{\lambda})\times(C_x+t-t_{j+1})}, \bar{\mathcal{L}}^n(B^n(t_j), \bar{B}^n(t_j,t_{j+1}))\rangle \\
\geq\ &[\tilde{B}(t_{j+1})-\tilde{B}(t_j)]\nu_F(C_0+\frac{\tilde{R}_{L,j}}{\lambda})\nu_G(C_x+t-t_j)-(\frac{3}{\lambda}+2)M_B\epsilon-3\epsilon.
\end{aligned}$$

Note that $\sum_{j=0}^{J-1}[\tilde{B}(t_{j+1})-\tilde{B}(t_j)]F^c(\frac{\tilde{R}_{U,j}}{\lambda})G^c(x+t-t_j)$ and $\sum_{j=0}^{J-1}[\tilde{B}(t_{j+1})-\tilde{B}(t_j)]F^c(\frac{\tilde{R}_{L,j}}{\lambda})G^c(x+t-t_{j+1})$ serve as the upper and lower Reimann sum of the integral $\int_{t_0}^t F^c(\frac{\tilde{R}(s)}{\lambda})G^c(x+t-s)d\tilde{B}(s)$, which converge to the integration as $n\to\infty$. So, by (10.41) and (10.42), we have that for all large n,

$$\left|\bar{\mathcal{Z}}^n(t)(C^\epsilon_x)-\bar{\mathcal{Z}}^n(0)(C^\epsilon_x+t)-\int_{t_0}^t F^c(\frac{\tilde{R}(s)}{\lambda})G^c(x+t-s)d\tilde{B}(s)\right|\leq(\frac{3}{\lambda}+2)M_B\epsilon+3\epsilon+\epsilon.$$

In summary, the right hand side of (10.52) can be bounded by a finite multiple of ϵ. We conclude that the left hand side of (10.52) must be 0 since it does not depend on ϵ, which can be arbitrary. This verifies (9.2).

The verification of fluid constrains (9.3) and (9.4) is quite straightforward. Basically, it is just passing the fluid scaled stochastic constraints

$$\bar{Q}^n(t) = (\bar{X}^n(t) - 1)^+,$$
$$\bar{Z}^n(t) = (\bar{X}^n(t) \wedge 1),$$

to $n \to \infty$. We omit it for brevity. □

APPENDIX A

A CONVOLUTION EQUATION

Lemma A.1. *Suppose $F(0) < 1$, $\rho > 0$ and $h(\cdot)$ is a càdlàg function. There exists a $b > 0$ (only depending on ρ and F) such that the two-side convolution equation* (3.23)

$$x(u) = h(u) + \int_0^u (x(u-v) - K)^+ dF(v) + \rho \int_0^u (x(u-v) \wedge K) dF_e(v).$$

has a unique solution $x(\cdot)$ on $[0, b]$. Furthermore, $x(\cdot)$ is càdlàg.

Proof. The space $\mathbf{D}([0, b], \mathbb{R})$ (all real valued *càdlàg* functions on $[0, b]$, c.f. Section 1.4) is a subset of the Banach space of bounded, measurable functions on $[0, b]$, equipped with the sup norm. One can check that this subset is closed in the Banach space. Thus, the space $\mathbf{D}([0, b], \mathbb{R})$ itself, equipped with the uniform metric υ_b (defined in Section 1.4), is complete.

Since $F(0) < 1$, there exists $b > 0$ such that

$$\kappa := \rho F_e(b) + F(b) < 1.$$

For any $y \in \mathbf{D}([0, b], \mathbb{R})$, define $\Psi(y)$ by

$$\Psi(y)(u) = h(u) + \rho \int_0^u (y(u-v) \wedge K) dF_e(v) + \int_0^u (y(u-v) - K)^+ dF(v),$$

for any $u \in [0, b]$. By convention, the integration $\int_0^u y(u-v) dF(v)$ is interpreted to be $\int_{(0,u]} y(u-v) dF(v)$ (c.f. Page 43 in [10]).

First, we show that Ψ is a mapping from $\mathbf{D}([0, b], \mathbb{R})$ to $\mathbf{D}([0, b], \mathbb{R})$. Since the function h is a *càdlàg* function, essentially we only need to show that the convolution $z(u) = \int_0^u y(u-v) dF(v)$ is a *càdlàg* function for any *càdlàg* function y and distribution function F. By Theorem 12.2.2 in [54], there exists a sequence of piece-wise constant

càdlàg functions y_n such that $\upsilon_b[y_n, y] \to 0$ as $n \to \infty$. By piece-wise constant *càdlàg*, we mean a function of the form

$$\sum_{j=0}^{J-1} c_j 1_{[a_j, b_j)} + c_J 1_{[a_J, b]},$$

where $c_j \in \mathbb{R}$, $a_j, b_j \in [0, b]$ with $a_j < b_j$ for all $j = 0, \ldots, J-1$ and $a_J < b$. Note that the convolution of indicator function $1_{[a_j, b_j)}$,

$$\int_0^u 1_{[a_j, b_j)}(u - v) dF(v)$$

equals 0 if $u < a_j$, equals $F(u - a_j) - F(0)$ if $u \in [a_j, b_j)$ and equals $F(u - a_j) - F(u - b_j)$ if $u \geq b_j$. Since F is *càdlàg*, the convolution of $1_{[a_i, b_j)}$ is also *càdlàg*. Similarly, the convolution of indicator function $1_{[a_J, b]}$,

$$\int_0^u 1_{[a_J, b]}(u - v) dF(v)$$

equals 0 if $u < a_J$ and equals $F(u - a_J) - F(0)$ if $u \in [a_J, b]$. Again, this convolution is a *càdlàg* function. It is now easy to see that $z_n(u) = \int_0^u y_n(u - v) dF(v)$ is a *càdlàg* function for each n since it is a linear combination of *càdlàg* functions. For any n, we have that

$$\begin{aligned} \upsilon_b[z_n, z] &\leq \sup_{u \in [0,b]} \int_0^u \big| y_n(u - v) - y(u - v) \big| dF(v) \\ &\leq \int_0^u \upsilon_b[y_n, y] dF(v) \leq F(u) \upsilon_b[y_n, y]. \end{aligned}$$

So $\upsilon_b[y_n, y] \to 0$ implies that $\upsilon_b[z_n, z] \to 0$. Since the space $\mathbf{D}([0, b], \mathbb{R})$ is complete under the uniform metric, the limit z is a *càdlàg* function.

Next, we show that the mapping Ψ is a contraction. For any $y, y' \in \mathbf{D}([0, b], \mathbb{R})$

we have that

$$\begin{aligned}
\upsilon_b[\Psi(y),\Psi(y')] &\le \sup_{u\in[0,b]} \rho \int_0^u \big|(y(u-v)\wedge K) - (y'(u-v)\wedge K)\big| dF_e(v) \\
&\quad + \sup_{u\in[0,b]} \int_0^u \big|(y(u-v)-K)^+ - (y'(u-v)-K)^+\big| dF(v) \\
&\le \rho \int_0^u \upsilon_b[y,y'] dF_e(v) + \int_0^u \upsilon_b[y,y'] dF(v) \\
&\le \kappa \upsilon_b[y,y'].
\end{aligned}$$

Since $\kappa < 1$, the mapping Ψ is a contraction.

By the contraction mapping theorem (c.f. Theorem 3.2 in [30]), Ψ has a unique fixed point x, i.e. $x = \psi(x)$. This implies that x is the unique solution to equation (3.23). □

Lemma A.2. *Assume the same condition as in Lemma A.1. Let $x(\cdot) \in \mathbf{D}([0,a],\mathbb{R})$ be the solution to equation (3.23) on some interval $[0,a]$ with $F(a) < 1$. If $h(\cdot)$ satisfies the following condition*

$$h(u) = (h(0)\wedge K)[1-G(u)] + (h(0)-K)^+[1-F(u)], \tag{A.1}$$

where $h(0) \ge 0$, $F(\cdot)$ is the same probability distribution function as in (3.23) and $G(\cdot)$ is a probability distribution function, then the function

$$\lambda \int_0^u (x(v)\wedge K) dv - (x(u)-K)^+$$

is non-decreasing in u on the interval $[0,a]$.

Proof. To simplify the notation, let $q(u) = (x(u)-K)^+$, $z(u) = x(u)\wedge K$ and

$$b(u) = \lambda \int_0^u z(v) dv - q(u) \tag{A.2}$$

for all $u \in [0,a]$. We need to show that $b(\cdot)$ is an non-decreasing function on the interval $[0,a]$. It follows from the definition of $F_e(\cdot)$ that $\rho \int_0^u z(u-v) dF_e(v) =$

$\lambda \int_0^u z(v)dv - \lambda \int_0^u z(v)F(u-v)dv$. Plugging it into (3.23) gives

$$\begin{aligned} x(u) = h(u) + \lambda \int_0^u z(v)dv \\ + \int_0^u q(u-v)dF(v) - \lambda \int_0^u z(v)F(u-v)dv. \end{aligned}$$

Applying Fubini's Theorem (c.f. Theorem 8.4 in [38]) to the last integral in the above, we have

$$\begin{aligned} \lambda \int_0^u z(v)F(u-v)dv &= \lambda \int_0^u \int_0^{u-v} z(v)dF(x)dv \\ &= \lambda \int_0^u \int_0^{u-x} z(v)dvdF(x). \end{aligned}$$

So we obtain

$$x(u) - \lambda \int_0^u z(v)dv = h(u) + \int_0^u \big[q(u-v) - \lambda \int_0^{u-v} z(u-v)dv\big]dF(v).$$

According to the definition of $b(\cdot)$ in (A.2), we have

$$b(u) = z(u) - h(u) + \int_0^u b(u-v)dF(v). \tag{A.3}$$

It now remains to use (A.2) and (A.3) to argue that $b(\cdot)$ is non-decreasing on the interval $[0, a]$, i.e. for any $u, u' \in [0, a] > 0$ with $u \leq u'$, we have $b(u) \leq b(u')$. Applying (A.3), we have

$$\begin{aligned} b(u') - b(u) &= z(u') - z(u) - [h(u') - h(u)] + \int_0^{u'} b(u'-v)dF(v) + \int_0^u b(u-v)dF(v) \\ &= z(u') - z(u) - [h(u') - h(u)] \\ &\quad + \int_u^{u'} b(u'-v)dF(v) + \int_0^u [b(u'-v) - b(u-v)]dF(v). \end{aligned}$$

Note that by condition (A.1), we have

$$\begin{aligned} -[h(u') - h(u)] &= -(h(0) \wedge K)[G(u) - G(u')] - (h(0) - K)^+[F(u) - F(u')] \\ &= (h(0) \wedge K)[G(u') - G(u)] - b(0)[F(u') - F(u)], \end{aligned}$$

where the last equation is due to (3.23) and (A.2). So

$$b(u') - b(u) = z(u') - z(u) + (h(0) \wedge K)[G(u') - G(u)] + \int_u^{u'} [b(u' - v) - b(0)]dF(v) + \int_0^u [b(u' - v) - b(u - v)]dF(v). \tag{A.4}$$

Since $b \in \mathbf{D}([0, a], \mathbb{R})$, according to Theorem 6.2.2 in the supplement of [54], it is bounded on the interval $[0, a]$. Let

$$b^* = \inf_{\{(u,u') \in [0,a] \times [0,a] : u \leq u'\}} b(u') - b(u).$$

If $z(u') < K$, then $q(u') = 0$. Thus, by (A.2),

$$b(u') - b(u) = \lambda \int_u^{u'} z(v)dv + q(u),$$

which is always non-negative; if $z(u') = K$, then $z(u') - z(u) \geq 0$. So it follows from (A.4) that

$$\begin{aligned} b(u') - b(u) &\geq \int_u^{u'} [b(u' - v) - b(0)]dF(v) + \int_0^u [b(u' - v) - b(u - v)]dF(v) \\ &\geq \int_0^{u'} b^* dF(v) = b^* F(u'). \end{aligned}$$

Summarizing both cases, we have

$$b(u') - b(u) \geq \min(0, b^* F(u'))$$

for all $u, u' \in [0, a] > 0$ with $u \leq u'$. Suppose that $b^* < 0$, taking the infimum on both sides over the set $\{(u, u') \in [0, a] \times [0, a] : u \leq u'\}$ gives $b^* \geq F(a)b^*$. This implies that $[1 - F(a)]b^* \geq 0$. Since $F(a) < 1$, it contradicts to that $b^* < 0$. So we must have $b^* \geq 0$, this implies that $b(\cdot)$ is non-decreasing on $[0, a]$. □

APPENDIX B

A KEY RENEWAL THEOREM WITH UNIFORM CONVERGENCE

Let $\mathscr{H}$ denote the set of non-increasing functions $h : [0, \infty) \to \mathbb{R}_+$ which are uniformly integrable. In this section, we show the following uniform convergence of the key renewal theorem.

Lemma B.1. *Assume that F is a non-lattice probability distribution function with finite mean β, and U is the associated renewal function. For each $\epsilon > 0$ there exists an $x_{F,\mathscr{H}}$ such that when $x \geq x_{F,\mathscr{H}}$,*

$$\sup_{h \in \mathscr{H}} |h * U(x) - \int_0^\infty h(y)dy| < \epsilon.$$

Proof. The gap between $h * U(x)$ and $\int_0^\infty h(y)dy$ can be written as

$$\int_x^\infty h(y)dy + \int_0^x h(x-y)d[U(y) - \frac{y}{\beta}]. \tag{B.1}$$

The first term in (B.1) converges to 0 uniformly for all $h \in \mathscr{H}$ by uniform integrability. For any $x > 0$ and $N \in \mathbb{N}$, let $N_x = \lfloor x - N \rfloor$. Then for all $x > N$ the second term equals $I_1^N(x) + I_2^N(x)$, with

$$I_1^N(x) = \int_0^{x-N_x} h(x-y)d[U(y) - \frac{y}{\beta}],$$
$$I_2^N(x) = \sum_{i=0}^{N_x-1} \int_{x-N_x+i}^{x-N_x+i+1} h(x-y)d[U(y) - \frac{y}{\beta}].$$

Then we have

$$I_1^N(x) \leq h(N_x) \int_0^{N+1} d[U(y) - \frac{y}{\beta}] \leq h(x-N)C_{N,F},$$

for some $C_{N,F}$ since $h(\cdot)$ is non-increasing. By uniform integrability, for any $\epsilon > 0$, there exist an $x_{\mathscr{H}}$ such that when $x \geq x_{\mathscr{H}}$,

$$I_1^n(x) \leq \epsilon C_{N,F},$$

where $C_{N,F}$ only depends on F and N. By Blackwell's renewal theorem (cf. Theorem 4.4 in Chapter 5 of [2]), for each $\epsilon > 0$ there exists an $N > 0$ such that $|U(x+1) - U(x) - \frac{1}{\beta}| < \epsilon$ whenever $x > N$. So

$$\begin{aligned} I_2^N(x) &\leq \epsilon \sum_{i=0}^{N_x-1} h(x - N_x + i) \\ &\leq \epsilon \sum_{i=0}^{N_x-1} \int_{x+N_x+i-1}^{x+N_x+i} h(y)dy \\ &= \epsilon \int_{x-N_x-1}^{x-1} h(y)dy \\ &\leq \epsilon \int_{N-2}^{x-1} h(y)dy, \end{aligned}$$

since $h(\cdot)$ is non-increasing and $N_x = \lfloor x_N \rfloor$. By uniform integrability, we can choose N big enough such that $\int_{N-2}^{x-1} h(y)dy \leq \epsilon$ for all $h \in \mathscr{H}$. Choose $x_{F,\mathscr{H}} = \max(N, x_{\mathscr{H}})$, then for all $x \geq x_{F,\mathscr{H}}$ we have

$$I_1^n(x) + I_2^N(x) \leq (C_{N,F} + \epsilon)\epsilon.$$

This completes the proof. □

APPENDIX C

SOME RESULTS ON THE PROHOROV METRIC

We prove some results on the Prohorov Metric. They are used in Chapter 5.

Lemma C.1. *Let μ and μ_1 be finite Borel measures on $[0,\infty)$. Denote $A_y = (y,\infty)$ for all $y \geq 0$. Let $M = \langle \chi, \mu \rangle$. For all $0 < \epsilon < 1$ if*

$$\sup_{y \geq 0} |\mu(A_y) - \mu_1(A_y)| < \epsilon, \tag{C.1}$$

then

$$\mathbf{d}[\mu, \mu_1] < (M+2)\epsilon^{1/3}.$$

Proof. Let α, β be positive constants to be determined later. Note that by Markov's inequality

$$\mu(\epsilon^{-\alpha}, \infty) \leq M\epsilon^{\alpha}.$$

For any real number a, denote $I_a = (a, a+\epsilon^{\beta}]$. Condition (C.1) implies that

$$\sup_{a \in \mathbb{R}_+} |\mu(I_a) - \mu_1(I_a)| < 2\epsilon.$$

For any Borel set $A \subset [0,\infty)$, there exist $a_1, \cdots, a_N$ such that

$$A \cap [0, \epsilon^{-\alpha}] \subset \cup_{i=1}^{N} I_{a_i},$$

and $I_{a_i} \cap I_{a_j} = \emptyset$ for all $i \neq j$, and $I_{a_i} \cap A \neq \emptyset$ for all i. These conditions imply that

$$N \leq \epsilon^{-\alpha-\beta}$$

and

$$\cup_{i=1}^{N} I_{a_i} \subset A^{\epsilon^{\beta}},$$

where A^{ϵ^β} is the ϵ^β-enlargement of the set defined in Section 1.4. So we have

$$\begin{aligned}\mu(A) &\leq \mu(A \cap [0, \epsilon^{-\alpha}]) + \mu(A \cap [\epsilon^{-\alpha}, \infty)) \\ &\leq \mu(\cup_{i=1}^N I_{a_i}) + M\epsilon^\alpha \\ &\leq \mu_1(\cup_{i=1}^N I_{a_i}) + N2\epsilon + M\epsilon^\alpha \\ &\leq \mu_1(A^{\epsilon^\beta}) + 2\epsilon^{1-\alpha-\beta} + M\epsilon^\alpha.\end{aligned}$$

Now choose $\alpha = \beta = 1/3$ to obtain

$$\mu(A) \leq \mu_1(A^{(M+2)\epsilon^{1/3}}) + (M+2)\epsilon^{1/3}.$$

The result of this lemma now follows from Property (ii) on page 72 in [5]. □

Lemma C.2. *Suppose μ_1 and μ are finite Borel measures on $\mathbb{R}_+$ satisfying*

$$\mathbf{d}[\mu_1, \mu] < \epsilon < 1, \tag{C.2}$$

and $\langle \chi^{1+q}, \mu_1 \rangle < M$, $\langle \chi^{1+q}, \mu \rangle < M$ for some positive constants q and M, then

$$|\langle \chi, \mu_1 \rangle - \langle \chi, \mu \rangle| \leq \epsilon^{1/2} + 2M\epsilon^{q/2}.$$

Proof. By Markov inequality, $\mu_1(A_x) \leq \frac{M}{x^{1+q}}$ and $\mu(A_x) \leq \frac{M}{x^{1+q}}$ for all $x \geq 0$. For any $C > 0$, we have the following inequality

$$\begin{aligned}&|\langle \chi, \mu_1 \rangle - \langle \chi, \mu \rangle| \\ &\leq \int_0^C |\mu_1(A_x) - \mu(A_x)| dx + \int_C^\infty \mu_1(A_x) dx + \int_C^\infty \mu(A_x) dx \\ &\leq C\epsilon + 2\int_C^\infty \frac{M}{x^{1+q}} dx = C\epsilon + 2M\frac{1}{C^q}.\end{aligned}$$

The result follows by letting $C = \epsilon^{-\frac{1}{2}}$. □

APPENDIX D

GLIVENKO-CANTELLI ESTIMATE

For any r, consider the sequence of i.i.d random variables $\{v_i^r\}_{i=-\infty}^{\infty}$ with law ν^r. In our setting, those v_i^r's with $i \geq 1$ correspond to the service requirement of the arriving jobs in the rth system; those with $i \leq 0$ correspond to the service requirement of initial jobs waiting in the buffer. For any $n \in \mathbb{Z}$ and $l \in \mathbb{R}_+$, define

$$\bar{\eta}^r(n,l) = \frac{1}{r} \sum_{i=n+1}^{n+\lfloor rl \rfloor} \delta_{v_i^r}. \tag{D.1}$$

The objective of this section is to obtain the *Glivenko-Cantelli Estimate*, Lemma D.2 below, for $\bar{\eta}^r(n,l)$. Very similar result was shown in Lemma 4.7 [24]. For completeness, the proof which follows the one in [24] is provided here.

To present the result, we introduce some notions from empirical process theory. Our primary references are [24] and [53].

A collection $\mathcal{C}$ of subsets of $\mathbb{R}^2$ shatters an n-point subset $\{x_1, \cdots, x_n\} \subset \mathbb{R}_+$ if the collection $\{\mathcal{C} \cap \{x_1, \cdots, x_n\} : C \in \mathcal{C}\}$ has cardinality 2^n. In this case, we say that $\mathcal{C}$ picks out all subsets of $\{x_1, \cdots, x_n\}$. The *Vapnik-Červonenkis index (VC-index)* of $\mathcal{C}$ is

$$V_{\mathcal{C}} = \min\{n : \mathcal{C} \text{ shatters no } n-\text{point subset}\},$$

where the minimum of the empty set equals infinity. The collection $\mathcal{C}$ is a *Vapnik-Červonenkis class (VC-class)* if it has finite VC-index. Let $\mathscr{V}$ be a family of Borel measurable functions $f : R_+ \to \mathbb{R}$. We call $\mathscr{V}$ a VC-class if the collection of the collection of subgraphs $\{\{(x,y) : y < f(x)\} : f \in \mathscr{V}\}$ is a VC-class of sets in $\mathbb{R}^2$.

VC-classes satisfy a very useful entropy bound. Let Γ be the set of all Borel probability measures γ on $\mathbb{R}_+$. for all $\gamma \in \Gamma$, denote $L_1(\gamma)$ the space of all Borel

measurable functions $f : \mathbb{R}_+ \to \mathbb{R}$ equipped with $L_1(\gamma)$-norm

$$\|f\|_{\gamma,1} = \langle |f|, \gamma \rangle.$$

For any $f \in L_1(\gamma)$, let $B_\gamma(f, \epsilon) = \{g \in \mathscr{V} : \|g - f\|_{\gamma,1} < \epsilon\}$ denote the $L_1(\gamma)$-ball in $L_1(\gamma)$, centered at f with radius ϵ. For a family of functions $\mathscr{V}$, $N(\epsilon, \mathscr{V}, L_1(\gamma))$ is the smallest number of balls $B_\gamma(f, \epsilon)$ needed to cover $\mathscr{V}$. Since $\mathscr{V}$ is the set of index functions over a VC-class $\mathcal{C}$,

$$\sup_{\gamma \in \Gamma} \log N(\epsilon \|\bar{f}\|_{\gamma,1}, \mathscr{V}, L_1(\gamma)) < \infty; \tag{D.2}$$

see Theorem 2.6.4 in [53].

We call a family of functions $\mathscr{V}$ a *Borel measurable class* if, for each $n \in \mathbb{N}$ and $(e_1, \cdots, e_n) \in \{-1, 1\}^n$, the map

$$(x_1, \cdots, x_n) \to \sup_{f \in \mathscr{V}} \sum_{i=1}^{n} e_i f(x_i)$$

is Borel measurable on $\mathbb{R}_+^n$. The condition requires that, for all $\delta > 0$ and $r \in \mathbb{R}_+$, the families $\mathscr{V}_\delta^r = \{f - g : f, g \in \mathscr{V}, \|f - g\|_{\nu^r,2} < \delta\}$ and $\mathscr{V}_\infty^2 = \{(f - g)^2 : f, g \in \mathscr{V}\}$ are Borel measurable, where

$$\|f\|_{\nu^r,2} = \langle |f|^2, \nu^2 \rangle,$$

denotes the $L_2(\nu^r)$-norm.

We call a Borel measurable function $\bar{f} : R_+ \to \mathbb{R}$ an envelope function for $\mathscr{V}$ if any element in $\mathscr{V}$ is bounded by $\bar{f}$. A VC-class with an envelop function satisfies a very useful entropy bound. Let Γ be the set of finitely discrete probability measures γ on $\mathbb{R}_+$ such that $\|\bar{f}\|_{\gamma,2} > 0$. For any Borel measurable function $f : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\|f\|_{\gamma,2} < \infty$, let $B_f(\epsilon) = \{g \in \mathscr{V} : \|g - f\|_{\nu^r,2} < \epsilon\}$ denote the $L_2(\nu^r)$-ball in $\mathscr{V}$, centered at f with radius ϵ. For a family of functions $\mathscr{V}$, $N(\epsilon, \mathscr{V}, L_2(\gamma))$ is the smallest number of balls $B_f(\epsilon)$ needed to cover $\mathscr{V}$. Then $\mathscr{V}$ satisfies

$$\int_0^\infty \sup_{\gamma \in \Gamma} \sqrt{\log N(\epsilon \|\bar{f}\|_{\gamma,2}, \mathscr{V}, L_2(\gamma))} d\epsilon < \infty; \tag{D.3}$$

see Definition 2.1.5, (2.5.1) and Theorem 2.6.7 in [53].

The Skorohod representation theorem implies the existence of $\mathbb{R}_+$-valued random variables $Y^r \sim \nu^r$ and $Y \sim \nu$ such that $Y^r \to Y$ almost surely. Thus there exists an $\mathbb{R}_+$-valued random variable $\bar{Y}$ such that

$$\bar{Y} = \sup_{r \in \mathbb{R}_+} Y^r, \quad \text{almost surely.} \tag{D.4}$$

Let $\bar{\nu}$ be the law of $\bar{Y}$. Since $L_2(\bar{\nu})$, the space of all Borel measurable functions $f : \mathbb{R}_+ \to \mathbb{R}$ equipped with $L_2(\bar{\nu})$-norm $\|f\|_{\bar{\nu},2} = \langle |f|^2, \bar{\nu} \rangle$, contains a continuous, increasing and unbounded function $\bar{f}$ such that

$$\mathbb{E}[\bar{f}(\bar{Y})^2] = \langle \bar{f}^2, \bar{\nu} \rangle < \infty. \tag{D.5}$$

Since $1_C \leq \bar{f}$ for all $C \in \mathcal{C}$, $\bar{f}$ is an envelope function for $\mathscr{V}$. Finally, denote $\overline{\mathscr{V}} = \mathscr{V} \cup \{\bar{f}\}$.

The objective of this section is to obtain the following *Glivenko-Cantelli Estimates* for $\bar{\eta}^r(n,l)$.

Lemma D.1. *Assume that* $\mathbf{d}[\nu^r, \nu] \to 0$ *as* $r \to \infty$*, where* ν *is a probability measure. Fix constants* $M_0, M_1, L > 0$*. For all* $\epsilon, \eta > 0$*,*

$$\limsup_{r \to \infty} \mathbb{P}\left(\max_{-rM_0 < n < rM_1} \sup_{l \in [0,L]} \sup_{f \in \overline{\mathscr{V}}} \left| \langle f, \bar{\eta}^r(n,l) \rangle - l \langle f, \nu^r \rangle \right| > \epsilon \right) < \eta. \tag{D.6}$$

Lemma D.2. *Let* $\mathscr{V}$ *be a VC-class of Borel measurable functions such that* $\mathscr{V}_\infty^2$ *and* $\mathscr{V}_\delta^r$ *are Borel measurable classes for all* $r \in \mathbb{R}_+$ *and* $\delta > 0$*. Assume there exists an envelop function* $\bar{f}$ *of* $\mathscr{V}$ *such that*

$$\lim_{N \to \infty} \sup_{r \in \mathbb{R}_+} \langle \bar{f} 1_{\bar{f} > N}, \nu^r \rangle = 0. \tag{D.7}$$

Fix constants $M_1, L_1 > 0$*. For all* $\epsilon, \epsilon' > 0$*,*

$$\limsup_{r \to \infty} \mathbb{P}\left(\max_{-rM_1 < n < r^2 M_1} \sup_{l \in [0,L_1]} \sup_{f \in \mathscr{V}} |\langle f, \bar{\eta}^r(n,l) \rangle - l \langle f, \nu^r(A_x) \rangle| > \epsilon' \right) < \epsilon. \tag{D.8}$$

Proof of Lemma D.1. Define

$$\bar{\eta}^r(l) = \frac{1}{r} \sum_{i=\lfloor -rM_0 \rfloor + 1}^{\lfloor -rM_0 \rfloor + \lfloor rl \rfloor} \delta_{v_i^r}.$$

By (D.1), it suffices to show that

$$\limsup_{r\to\infty} \mathbb{P}\left(\sup_{l\in[0,L']} \sup_{f\in\overline{\mathscr{V}}} \left| \langle f, \bar{\eta}^r(l) \rangle - l\langle f, \nu^r \rangle \right| > \epsilon/2 \right) < \eta, \tag{D.9}$$

where $L' = L + M_0 + M_1$.

We now apply Theorem 2.8.1 in [53] to show (D.9). Observe that for all $n \in \mathbb{N}$ and $(e_1, \cdots, e_n) \in \mathbb{R}^n$, the function

$$(x_1, \cdots, x_n) \to \sup_{f\in\overline{\mathscr{V}}} \sum_{i=1}^{n} e_i f(x_i)$$

is measurable on the completion of $(\bar{\mathbb{R}}_+^2, \mathscr{B}, \nu^r)^n$, for all $r \in \mathbb{R}_+$. Thus $\overline{\mathscr{V}}$ a ν^r-measurable class for all $r \in \mathbb{R}_+$; see Definition 2.3.3 in [53]. Moreover, $\overline{\mathscr{V}}$ is uniformly bounded above by the envelope function $\bar{f}$ and

$$\lim_{M\to\infty} \sup_{r\in\mathbb{R}_+} \langle \bar{f} 1_{\{\bar{f}>M\}}, \nu^r \rangle = 0, \tag{D.10}$$

by Markov's inequality, (D.4) and (D.5). Lastly, $\overline{\mathscr{V}}$ satisfies the finite entropy bound (D.2) because $N(\epsilon, \overline{\mathscr{V}}, L_1(\gamma)) \leq N(\epsilon, \mathscr{V}, L_1(\gamma)) + 1$ and $\mathcal{C}$ is a VC-class. These three observations imply that the assumptions of Theorem 2.8.1 in [53] are satisfied. Consequently, $\overline{\mathscr{V}}$ is *Glivenko-Cantelli*, uniformly in r. That is for every $\delta > 0$, there exists an n_δ such that

$$\limsup_{r\to\infty} \mathbb{P}\left(\sup_{m\geq n_\delta} \sup_{f\in\overline{\mathscr{V}}} \left| \frac{1}{m} \sum_{-rM_0+1}^{-rM_0+m} f(v_i^r) - \langle f, \nu^r \rangle \right| > \delta \right) < \delta. \tag{D.11}$$

Note that the probability on the left side of (D.9) can be upper bounded by

$$\mathbb{P}\left(\sup_{l\in[0,L']} \sup_{f\in\overline{\mathscr{V}}} \frac{\lfloor rl \rfloor}{r} \left| \frac{1}{\lfloor rl \rfloor} \sum_{-rM_0+1}^{-rM_0+\lfloor rl \rfloor} f(v_i^r) - \langle f, \nu^r \rangle \right| > \epsilon/4 \right)$$
$$+ \mathbb{P}\left(\frac{1}{r} \sup_{f\in\overline{\mathscr{V}}} \langle f, \nu^r \rangle > \frac{\epsilon}{4} \right).$$

By (D.4) and (D.5), the second term in the above vanishes as $r \to \infty$. The first term can be upper bounded by

$$\begin{aligned} &\mathbb{P}\left(\frac{n_\delta}{r} \sup_{m\in[0,n_\delta]} \sup_{f\in\overline{\mathscr{V}}} \Big|\frac{1}{m} \sum_{-rM_0+1}^{-rM_0+m} f(v_i^r) - \langle f, \nu^r\rangle\Big| > \epsilon/4\right) \\ &+ \mathbb{P}\left(L \sup_{m\in[n_\delta,L'r]} \sup_{f\in\overline{\mathscr{V}}} \Big|\frac{1}{m} \sum_{-rM_0+1}^{-rM_0+m} f(v_i^r) - \langle f, \nu^r\rangle\Big| > \epsilon/4\right) \end{aligned} \tag{D.12}$$

To see this, one can replace m by $\lfloor rl \rfloor$ and devide the interval $[0, L']$ into $[0, n_\delta/r]$ and $[n_\delta, L']$. Denote

$$X(f) = \sup_{m\in[0,n_\delta]} \Big|\frac{1}{m} \sum_{-rM_0+1}^{-rM_0+m} f(v_i^r) - \langle f, \nu^r\rangle\Big|.$$

When $f \in \mathscr{V}$, it is clear that $X(f) \leq 2$. By (D.4) and (D.5), $X(\bar{f})$ is a random variable with finite mean and variance. So there exists a constant M_3 such that

$$\mathbb{P}\left(\sup_{f\in\overline{\mathscr{V}}} X(f) > M_3\right) < \eta/2.$$

So the first term in (D.12) is bounded by $\eta/2$ for all $r \geq 4M_3 n_\delta/\epsilon$. According to (D.11), the lim sup of the second term in (D.12) will be bounded by $\eta/2$ if we choose $\delta = \min(\frac{\epsilon}{4L}, \eta/2)$. □

To better structure the proof of the second result, we present the following auxiliary lemma.

Lemma D.3. *For $n \in \mathbb{Z}$ and $k \in \mathbb{N}$, define*

$$\xi_{n,k}^r = \frac{1}{\sqrt{k}} \sum_{i=n+1}^{n+k} (\delta_{v_i^r} - \nu^r). \tag{D.13}$$

Then for any $q > 1$, $y > 2$ and $n \in \mathbb{Z}$ there exists $M_q < \infty$ and k_0 such that $k \geq k_0$ implies

$$\sup_r \mathbb{P}\left(\sup_{f\in\mathscr{V}} \langle f, \xi_k^r\rangle > y\right) < \frac{M_q}{y^q}. \tag{D.14}$$

The constant M_q does not depend on y.

Proof. Let us first fix $n = 0$ and look at $\xi^r_{0,k}$ which will be denoted by ξ^r_k for simplicity. The property of the envelop function $\bar{f}$ (D.7) and the uniform entropy bound (D.3), together with the sets $\mathscr{V}^r_\delta$ and $\mathscr{V}^r_\infty$ to be Borel measurable, imply that $\mathscr{V}$ is Donsker and pre-Gaussian uniformly in ν^r, $r \in \mathbb{R}_+$. (See Theorem 2.8.3 in [53].)

Let $l^\infty(\mathscr{V})$ be the space of all probability measures on $\mathbb{R}_+$ equipped with norm $\|\cdot\|_\mathscr{V} = \sup_{f\in\mathscr{V}}\langle f, \cdot\rangle$. $\mathscr{V}$ being Donsker uniformly in ν^r means that ξ^r_k converges weakly as $n \to \infty$ in $l^\infty(\mathscr{V})$ to a tight, Borel measurable version of the Brownian bridge ξ^r uniformly for all ν^r. According to Chapter 1.12 in [53], this is equivalent to

$$\sup_{h\in \mathrm{BL}_1} |\mathbb{E}^r h(\xi^r_k) - \mathbb{E}h(\xi^r)| \to 0. \tag{D.15}$$

uniformly for all ν^r, where BL_1 is the set of functions $h : l^\infty(\mathscr{V}) \to \mathbb{R}$ which are uniformly bounded by 1 and satisfy $|h(z_1) - h(z_2)| \le \|z_1 - z_2\|_\mathscr{V}$. Pre-Gaussian uniformly in ν^r means that

$$\sup_r \mathbb{E}^r[\sup_{f\in\mathscr{V}}\langle f, \xi^r\rangle] < \infty. \tag{D.16}$$

Define $h_y : l^\infty(\mathscr{V}) \to \mathbb{R}$ by

$$h_y(\cdot) = (\sup_{f\in\mathscr{V}}\langle f, \cdot\rangle - y + 1)^+ \wedge 1.$$

Then it is clear that $h_y \in \mathrm{BL}_1$, and

$$\sup_r \mathbb{P}\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r_k\rangle > y\right) \le \sup_r \mathbb{E}^r\Big[h_y(\xi^r_k)\Big].$$

By (D.15) and the above inequality, there exists $k_0 \in \mathbb{N}$ such that $k \ge k_0$ implies

$$\sup_r \mathbb{P}\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r_k\rangle > y\right) \le \sup_r \mathbb{E}^r\Big[h_y(\xi^r)\Big] + y^{-q}.$$

Applying the definition of h_y and Markov inequality to obtain

$$\begin{aligned}\sup_r \mathbb{P}\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r_k\rangle > y\right) &\le \sup_r \mathbb{P}\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r\rangle > y - 1\right) + y^{-q} \\ &\le y^{-q}\Big(2^q \sup_r \mathbb{E}^r\big[\sup_{f\in\mathscr{V}}\langle f, \xi^r\rangle\big]^q + 1\Big).\end{aligned}$$

Let M_q be the last term in parentheses, which does not depend on y. For each $r \in \mathbb{R}_+$, the Brownian bridge is separable and Gaussian with $\sup_{f\in\mathscr{V}}\langle f, \xi^r\rangle$ finite almost surely. Thus, there exist a constant M such that for all $r \in \mathbb{R}_+$,

$$\mathbb{E}^r\big[\sup_{f\in\mathscr{V}}\langle f, \xi^r\rangle\big]^q \le M\big[\mathbb{E}^r \sup_{f\in\mathscr{V}}\langle f, \xi^r\rangle\big]^q,$$

see Proposition A.2.4 in [53]. Conclude from (D.16) that $M_q < \infty$.

So far, we have shown that the result (D.14) is true for $n = 0$. Note that for any $n \in \mathbb{Z}$, $\xi^r_{n,k}$ is defined on the shifted sequence $v^r_{n+1}, v^r_{n+2}, \cdots$ By the i.i.d property of the sequence, if we fix k then $\xi^r_{n,k}$ has the same distribution for all $n \in \mathbb{Z}$. So we can conclude that (D.14) is true for all $n \in \mathbb{Z}$. □

Proof of Lemma D.2. Note that

$$|\langle f, \bar{\eta}^r(n, l)\rangle - l\langle f, \nu^r\rangle| \le \frac{1}{r}\sum_{i=n+1}^{n+\lfloor rl\rfloor}[\langle f, \delta_{v^r_i}\rangle - \langle f, \nu^r\rangle] + \frac{1}{r}.$$

Since for each $\epsilon' > 0$, $1/r < \epsilon'/2$ for all large r, so the probability in (D.8) can be bounded by

$$\limsup_{r\to\infty}\mathbb{P}\left(\max_{-rM_1<n<r^2M_1}\sup_{l\in[0,L]}\sup_{f\in\mathscr{V}}|\frac{1}{r}\sum_{i=n+1}^{n+\lfloor rl\rfloor}[\langle f, \delta_{v^r_i}\rangle - \langle f, \nu^r\rangle]| > \frac{\epsilon'}{2}\right). \tag{D.17}$$

Pick $\delta > 0$, when r is large enough ($r > M_1/\delta$) the interval $[-rM_1, r^2M_1]$ will be covered by intervals

$$[-r^2\delta, 0], [0, r^2\delta], \cdots, [(\lceil\frac{M_1}{\delta}\rceil - 1)r^2\delta, \lceil\frac{M_1}{\delta}\rceil r^2\delta].$$

When r is large enough ($r^2\delta > \lfloor rL_1\rfloor$), (D.17) can be further bounded by

$$\limsup_{r\to\infty}\mathbb{P}\left(\max_{-1\le j\le\lceil\frac{M_1}{\delta}\rceil-1}\max_{0\le k,k'\le r^2\delta}\sup_{f\in\mathscr{V}}|\frac{\sqrt{k}}{r}\langle f, \xi^r_{jr^2\delta,k}\rangle - \frac{\sqrt{k'}}{r}\langle f, \xi^r_{jr^2\delta,k'}\rangle| > \frac{\epsilon'}{2}\right).$$

Since $\xi^r_{n,\cdot}$ has stationary increments, the previous term can be bounded above by

$$\limsup_{r\to\infty}\lceil\frac{M_1}{\delta}\rceil\mathbb{P}\left(\max_{0\le k\le r^2\delta}\sup_{f\in\mathscr{V}}|\frac{\sqrt{k}}{r}\langle f, \xi^r_{0,k}\rangle| > \frac{\epsilon'}{2}\right).$$

By Ottaviani's inequality (see Proposition $A.1.1$ in [53]) and by stationary increments of $\xi^r_{n,\cdot}$, this can be bounded above by

$$\limsup_{r\to\infty} \frac{\lceil \frac{M_1}{\delta} \rceil \mathbb{P}\left(\sup_{f\in\mathscr{V}} \langle f, \xi^r_{0,\lfloor r^2\delta \rfloor}\rangle > \frac{\epsilon'}{4\sqrt{\delta}}\right)}{1 - \max_{0\le k\le r^2\delta} \mathbb{P}\left(\sup_{f\in\mathscr{V}} \langle f, \xi^r_{0,k}\rangle > \frac{\epsilon' r}{4\sqrt{k}}\right)}. \tag{D.18}$$

Assume δ is small enough so that $\frac{\epsilon'}{4\sqrt{\delta}} > 2$. By Lemma D.3, there exists M_3 and $k_0 \in \mathbb{N}$ such that $k > k_0$ implies

$$\sup_r \mathbb{P}\left(\sup_{f\in\mathscr{V}} \langle f, \xi^r_{0,k}\rangle > \frac{\epsilon'}{4\sqrt{\delta}}\right) \le \left(\frac{4\sqrt{\delta}}{\epsilon'}\right)^3 M_3.$$

Since $\lfloor r^2\delta \rfloor \to \infty$ as $r \to \infty$, the limit superior of the numerator in (D.18) can be bounded above by $\lceil M_1/\delta \rceil (4\sqrt{\delta}/\epsilon')^3 M_3$, which can be made arbitrarily small by choosing δ sufficiently small. By the same reason, those terms in the maximum of the denominator with index $k > k_0$ are bounded above by $(4\sqrt{\delta}/\epsilon')^3 M_3$. For those terms with index $k \le k_0$,

$$\mathbb{P}\left(\sup_{f\in\mathscr{V}} \langle f, \xi^r_{0,k}\rangle > \frac{\epsilon' r}{4\sqrt{k}}\right) \le \mathbb{P}\left(\sup_{f\in\mathscr{V}} \langle f, \xi^r_{0,k}\rangle > \frac{\epsilon' r}{4\sqrt{k_0}}\right),$$

which converges to zero as $r \to \infty$. By choosing δ small enough, (D.18) can be made arbitrarily small for all large r. □

APPENDIX E

ANOTHER CONVOLUTION EQUATION

Lemma E.1. *Assume that $G(\cdot)$ is a distribution function with $G(0) < 1$, $\zeta(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$, $H(\cdot)$ is a Lipschitz continuous function, and $\rho \in \mathbb{R}$. There exists a unique solution $x^*(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$ to the following equation:*

$$x(t) = \zeta(t) + \rho \int_0^t H\big((x(t-s)-1)^+\big) dG_e(s) + \int_0^t (x(t-s)-1)^+ dG(s), \qquad \text{(E.1)}$$

where, G_e is the equilibrium distribution of G.

Proof. Suppose $H(\cdot)$ is Lipschitz continuous with constant L. The equilibruim distribution has density $\mu[1-G(\cdot)]$, so $|G_e(t) - G_e(s)| \le \mu|t-s|$ for any $s,t \in \mathbb{R}$. Since $G(0) < 1$, there exists $b > 0$ such that

$$\kappa := \rho L[G_e(b) - G_e(0)] + [G(b) - G(0)] < 1.$$

Now consider the space $\mathbf{D}([0,b],\mathbb{R})$ (all real valued *càdlàg* functions on $[0,b]$) is a subset of the Banach space of bounded, measurable functions on $[0,b]$, equipped with the sup norm. One can check that this subset is closed in the Banach space. Thus, the space $\mathbf{D}([0,b],\mathbb{R})$ itself, equipped with the uniform metric υ_T, is complete.

For any $y \in \mathbf{D}([0,b],\mathbb{R})$, define $\Psi(y)$ by

$$\Psi(y)(t) = \zeta(t) + \rho \int_0^t H\left((y(t-s)-1)^+\right) dG_e(s) + \int_0^t (y(t-s)-1)^+ dG(s),$$

for any $t \in [0,b]$. By convention, the integration $\int_0^t y(t-s)dF(s)$ is interpreted to be $\int_{(0,t]} y(t-s)dF(s)$ (c.f. Page 43 in [10]). We prove the existence and uniqueness of the solution to equation (E.1) by showing that Ψ is a contraction mapping on $\mathbf{D}([0,b],\mathbb{R})$. According to the proof of Lemma A.1 in [63], the convolution of a *càdlàg*

function with a distribution function is still a *càdlàg* function. So Ψ is a mapping from $\mathbf{D}([0,b],\mathbb{R})$ to $\mathbf{D}([0,b],\mathbb{R})$. Next, we show that the mapping Ψ is a contraction. For any $y, y' \in \mathbf{D}([0,b],\mathbb{R})$, we have that

$$\begin{aligned}
\upsilon_b[\Psi(y),\Psi(y')] &\leq \sup_{t\in[0,b]} \rho \int_0^t L\big|(y(t-s)-1)^+ - (y'(u-v)-1)^+\big| dG_e(s) \\
&\quad + \sup_{t\in[0,b]} \int_0^t \big|(y(t-s)-1)^+ - (y'(t-s)-1)^+\big| dG(s) \\
&\leq \rho L \int_0^b \upsilon_b[y,y'] dG_e(s) + \int_0^b \upsilon_b[y,y'] dG(s) \\
&\leq \kappa \upsilon_b[y,y'].
\end{aligned}$$

Since $\kappa < 1$, the mapping Ψ is a contraction. By the contraction mapping theorem (c.f. Theorem 3.2 in [30]), Ψ has a unique fixed point x, i.e. $x = \psi(x)$. This implies that $x \in \mathbf{D}([0,b],\mathbb{R})$ is the unique solution to equation (E.1) on $[0,b]$.

It now remains to extend the existence and uniqueness result from $[0,b]$ to $[0,T]$. Denote $x_b(t) = x(b+t)$, $\zeta_b(t) = \zeta(b+t) + \rho \int_t^{b+t} H\left((x(b+t-s)-1)^+\right) dG_e(s) + \int_t^{b+t} (x(b+t-s)-1)^+ dG(s)$, then we have for $t \in [0, T-b]$,

$$x_b(t) = \zeta_b(t) + \rho \int_0^t H\left((x_b(t-s)-1)^+\right) dG_e(s) + \int_0^t (x_b(t-s)-1)^+ dG(s). \quad \text{(E.2)}$$

It follows from the previous argument that there is unique solution $x_b(\cdot)$ to the above equation. Thus, we obtain a unique extension of the solution to (E.1) on the interval $[0,2b]$. Repeating this approach for N time with $N \geq \lceil T/b \rceil$ gives a unique solution on the interval $[0,T]$. □

APPENDIX F

THE SPECIAL CASE WITH EXPONENTIAL DISTRIBUTIONS FOR MULTI-SERVER QUEUES

We verify here that the fluid model developed in Chapter 9 for the general patience and service time distributions is consistent with the one in [55], that was obtained in the special case where both distributions are assumed to be exponential.

Our fluid model equations implies the key relationship (9.21). Now, we specialize in the case with exponential distribution, i.e.

$$F(t) = F_e(t) = 1 - e^{-\alpha t}, \quad G(t) = G_e(t) = 1 - e^{-\mu t}, \quad \text{for all } t \geq 0.$$

Now (9.21) becomes

$$\bar{X}(t) = \zeta_0(t) + \rho \int_0^t \big[1 - \frac{\alpha}{\lambda}\big((\bar{X}(t-s) - 1)^+\big)\big]\mu e^{-\mu s} ds + \int_0^t (\bar{X}(t-s) - 1)^+ \mu e^{-\mu s} ds.$$

In the case of exponential service time distribution, the remaining service time of those initially in service and the service times of those initially waiting in queue are also assumed to be exponentially distributed. So we have

$$\zeta_0(t) = \bar{\mathcal{Z}}_0(C_0 + t) + \bar{Q}_0 e^{-\mu t} = \bar{X}_0 e^{-\mu t},$$

where $\bar{X}_0 = \bar{Z}_0 + \bar{Q}_0$ is the initial number of customers in the system. By some algebra, the above two equations can be simplified as the following,

$$\bar{X}(t) = \bar{X}_0 e^{-\mu t} + \rho[1 - e^{-\mu t}] + (\mu - \alpha) \int_0^t (\bar{X}(t-s) - 1)^+ e^{-\mu s} ds. \tag{F.1}$$

By the change of variable $t - s \to s$, the above integration can be written as

$$\int_0^t (\bar{X}(t-s) - 1)^+ e^{-\mu s} ds = e^{-\mu t} \int_0^t (\bar{X}(s) - 1)^+ e^{\mu s} ds.$$

Taking the derivative on both sides of (F.1) yields

$$\begin{aligned}
\bar{X}'(t) &= -\mu X_0 e^{-\mu t} + \mu\rho e^{\mu t} \\
&\quad + (\mu-\alpha)[-\mu e^{-\mu t}\int_0^t (\bar{X}(s)-1)^+ e^{\mu s} ds + e^{-\mu t}(\bar{X}(t)-1)^+ e^{\mu t}] \\
&= -\mu X_0 e^{-\mu t} - \mu\rho[1-e^{\mu t}] + \mu\rho \\
&\quad - \mu(\mu-\alpha)e^{-\mu t}\int_0^t (\bar{X}(s)-1)^+ e^{\mu s} ds + (\mu-\alpha)(\bar{X}(t)-1)^+ \\
&= -\mu\bar{X}(t) + \mu\rho + (\mu-\alpha)(\bar{X}(t)-1)^+.
\end{aligned}$$

Using the notation in [55], $a^- = -\min(0,a)$ for any $a \in \mathbb{R}$. Note that $a = \min(a,1) + (a-1)^+ = 1 - (a-1)^- + (a-1)^+$. So the above equation further implies

$$\bar{X}'(t) = \mu(\rho-1) - \alpha(\bar{X}(t)-1)^+ + \mu(\bar{X}(t)-1)^-, \quad \text{for all } t \geq 0.$$

This equation is consistent with Theorem 2.2 in [55] (μ is assumed to be 1 in that paper).

F.1 An Extension of Skorohod Representation Theorem

In this section, we present a slight extension, Lemma F.1 below, of the Skorohod Representation Theorem (c.f. Theorem 3.2.2 in [54]). The proof of Lemma F.1 is built on the proof of Theorem 3.2.2 provided in the supplement of [54], with slight extension to deal with the product of two matric spaces.

Let $(\mathbf{E}_1, \pi_1)$ and $(\mathbf{E}_2, \pi_2)$ be two complete and separable metric spaces. Let $(\mathbf{E}_1 \times \mathbf{E}_2, \pi)$ denote the product space of them, with the product metric π obtained by the maximum metric.

Lemma F.1. *Consider a sequence of random variables $\{(X_n, Y_n), n \geq 1\}$ in the product space $\mathbf{E}_1 \times \mathbf{E}_2$. If $X_n \Rightarrow X$, then there exists other random elements of $\mathbf{E}_1 \times \mathbf{E}_2$, $\{(\tilde{X}_n, \tilde{Y}_n), n \geq 1\}$, and $\tilde{X}$, defined on a common underlying probability space, such that*

$$(\tilde{X}_n, \tilde{Y}_n) \stackrel{d}{=} (X_n, Y_n), n \geq 1, \quad \tilde{X} \stackrel{d}{=} X$$

and almost surely,

$$\tilde{X}_n \to \tilde{X} \quad as\ n \to \infty.$$

Proof. In order to present the proof, we first need some preliminaries. A nested family of countably partitions of a set A is a collection of subsets $A_{i_1,\ldots,i_k}$ indexed by k-tuples of positive integers such that $\{A_i : i \geq 1\}$ is a partition of A and $\{A_{i_1,\ldots,i_{k+1}} : i_{k+1} \geq 1\}$ is a partition of $A_{i_1,\ldots,i_k}$ for all $k \geq 1$ and $(i_1,\ldots,i_k) \in \mathbb{N}_+^k$. Let $\mathbb{P}_1$ denote the probability measure on the space where X lives on. Since the space $(\mathbf{E}_1, \pi_1)$ is separable, according to Lemma 1.9 in the supplement of [54], there exists a nested family of countably partitions $\{E^1_{i_1,\ldots,i_k}\}$ of $(\mathbf{E}_1, \pi_1)$ that satisfies

$$\mathrm{rad}(E^1_{i_1,\ldots,i_k}) < 2^{-k}, \tag{F.2}$$

$$\mathbb{P}_1(\partial E^1_{i_1,\ldots,i_k}) = 0, \tag{F.3}$$

where $\mathrm{rad}(A)$ denotes the radius of the set A in a metric space, and $\partial(A)$ denote the boundary of the set A. Since the space $(\mathbf{E}_2, \pi_2)$ is separable, by the same lemma, there exists a nested sequence of countably partitions $\{E^2_{i'_1,\ldots,i'_{k'}}\}$ of $(\mathbf{E}_2, \pi_2)$ that satisfies

$$\mathrm{rad}(E^2_{i'_1,\ldots,i'_{k'}}) < 2^{-k'}. \tag{F.4}$$

Note that for space $(\mathbf{E}_2, \pi_2)$, we only need a weaker version of Lemma 1.9 in the supplement of [54].

The first step is to use this nested sequence of countably partitions to construct random variables $\{(\tilde{X}_n, \tilde{Y}_n), n \geq 1\}$ with the same distribution for each n. For $n \geq 1$, we first construct subintervals $I^n_{i_1,\ldots,i_k} \subseteq [0,1)$ corresponding to the marginal probability of X_n. Let $I^n_1 = [0, \mathbb{P}^n(E^1_1 \times \mathbf{E}_2))$ and

$$I^n_i = \Big[\sum_{j=1}^{i-1} \mathbb{P}^n(E^1_j \times \mathbf{E}_2), \sum_{j=1}^{i} \mathbb{P}^n(E^1_j \times \mathbf{E}_2)\Big), \quad i > 1,$$

where $\mathbb{P}^n$ is the probability measure on the space where (X_n, Y_n) lives. Let $\{I^n_{i_1,\ldots,i_{k+1}} :$

$i_{k+1} \geq 1\}$ be a countable partition of subintervals of $I^n_{i_1,\ldots,i_k}$. If $I^n_{i_1,\ldots,i_k} = [a_n, b_n)$, then

$$I^n_{i_1,\ldots,i_{k+1}} = \Big[a_n + \sum_{j=1}^{i_{k+1}-1} \mathbb{P}^n(E^1_{i_1,\ldots,i_k,j} \times \mathbf{E}_2), a_n + \sum_{j=1}^{i_{k+1}} \mathbb{P}^n(E^1_{i_1,\ldots,i_k,j} \times \mathbf{E}_2)\Big).$$

The length of each subinterval $I^n_{i_1,\ldots,i_k}$ is the probability $\mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times \mathbf{E}_2)$. We then construct further subintervals $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}} \subseteq I^n_{i_1,\ldots,i_k}$ corresponding to (X_n, Y_n). If $I^n_{i_1,\ldots,i_k} = [a_n, b_n)$, then let $I^n_{i_1,\ldots,i_k;1} = [a_n, a_n + \mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_1))$ and

$$I^n_{i_1,\ldots,i_k;i'} = \Big[a_n + \sum_{j'=1}^{i'-1} \mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_{j'}), a_n + \sum_{j'=1}^{i'} \mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_{j'})\Big), \quad i' > 1.$$

Let $\{I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'+1}} : i'_{k'+1} \geq 1\}$ be countable partition of $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}}$. If $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}} = [a_n, b_n)$, then

$$\begin{aligned} &I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'+1}} \\ &= \Big[a_n + \sum_{j'=1}^{i'_{k'+1}-1} \mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_{i'_1,\ldots,i'_k,j'}), a_n + \sum_{j'=1}^{i'_{k'+1}} \mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_{i'_1,\ldots,i'_k,j'})\Big). \end{aligned}$$

The length of each subinterval $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}}$ is the probability $\mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_{i'_1,\ldots,i'_{k'}})$. Now from each nonempty subset $E^1_{i_1,\ldots,i_k} \times E^2_{i'_1,\ldots,i'_k}$ we choose one point $(x_{i_1,\ldots,i_k}, y_{i'_1,\ldots,i'_k})$. For each $n \geq 1$ and $k \geq 1$, we define functions $(x^k_n, y^k_n) : [0,1) \to \mathbf{E}_1 \times \mathbf{E}_2$ by letting $x^k_n(w) = x_{i_1,\ldots,i_k}$ and $y^k_n(w) = y_{i'_1,\ldots,i'_k}$ for $\omega \in I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_k}$. By the nested partition property and inequalities F.2 and F.4,

$$\pi\big((x^k_n(\omega), x^k_n(\omega)), (x^{k+j}_n(\omega), x^{k+j}_n(\omega))\big) < 2^{-k} \quad \text{for all } j, k, n$$

and $\omega \in [0,1)$. Since $(\mathbf{E}_1 \times \mathbf{E}_2, \pi)$ is a complete metric space, the above implies that there is $(x_n(\omega), y_n(\omega)) \in \mathbf{E}_1 \times \mathbf{E}_2$ such that

$$\pi\big((x^k_n(\omega), x^k_n(\omega)), (x_n(\omega), x_n(\omega))\big) \to 0 \quad \text{as } k \to \infty.$$

We let $(\tilde{X}_n, \tilde{Y}_n) = (x_n, y_n)$ on $[0,1)$ for $n \geq 0$.

The next step is to construct $\tilde{X}$ and show that $\tilde{X}_n \to \tilde{X}$ almost surely. For each $n \geq 1$, let $\mathbb{P}^n_1$ denote the marginal probability of X^n. It is clear that $I^n_{i_1,\ldots,i_k}$ is

the probability $\mathbb{P}_1^n(E^1_{i_1,\dots,i_k})$. By (F.3), we have that $\mathbb{P}_1^n(E^1_{i_1,\dots,i_k}) \to \mathbb{P}_1(E^1_{i_1,\dots,i_k})$, as $n \to \infty$. Consequently, the length of the interval $I^n_{i_1,\dots,i_k}$ converges to the length of the interval $I_{i_1,\dots,i_k}$, which is defined in a similar way as for $I^n_{i_1,\dots,i_k}$ by letting

$$I_{i_1,\dots,i_{k+1}} = \Big[a_n + \sum_{j=1}^{i_{k+1}-1} \mathbb{P}_1(E_{i_1,\dots,i_k,j}), a_n + \sum_{j=1}^{i_{k+1}} \mathbb{P}_1(E_{i_1,\dots,i_k,j})\Big),$$

if $I_{i_1,\dots,i_k} = [a_n, b_n)$. Now from each nonempty subset $E_{i_1,\dots,i_k}$ we choose one point $x_{i_1,\dots,i_k}$. For each $k \geq 1$, we define functions $x^k : [0,1) \to \mathbf{E}_1$ by letting $x^k(\omega) = x_{i_1,\dots,i_k}$ for $\omega \in I^n_{i_1,\dots,i_k}$. By the nested partition property and inequalities F.2,

$$\pi_1(x^k(\omega), x^{k+j}(\omega)) < 2^{-k} \quad \text{for all } j,k$$

and $\omega \in [0,1)$. Since $(\mathbf{E}_1, \pi_1)$ is a complete metric space, the above implies that there is $x(\omega) \in \mathbf{E}_1$ such that

$$\pi_1(x^k(\omega), x(\omega)) \to 0 \quad \text{as } k \to \infty.$$

We let $\tilde{X} = x$ on $[0,1)$. Since

$$\begin{aligned}\pi_1(\tilde{X}_n(\omega), \tilde{X}(\omega)) &\leq \pi_1(\tilde{X}_n(\omega), \tilde{X}^k_n(\omega)) + \pi_1(\tilde{X}^k_n(\omega), \tilde{X}^k(\omega)) + \pi_1(\tilde{X}^k(\omega), \tilde{X}(\omega)) \\ &\leq 3 \times 2^{-k},\end{aligned}$$

for all ω in the interior of $I_{i_1,\dots,i_k}$,

$$\lim_{n\to\infty} \pi_1(\tilde{X}_n(\omega), \tilde{X}(\omega)) \leq 3 \times 2^{-k}.$$

Since k is arbitrary, we must have $\tilde{X}_n(\omega) \to \tilde{X}(\omega)$ as $n \to \infty$ for all but at most countably many $\omega \in [0,1)$.

It remains to show that $(\tilde{X}_n, \tilde{Y}_n)$ has the probability laws $\mathbb{P}^n$. Let $\tilde{\mathbb{P}}$ denote the Lebesque measure on $[0,1)$. It suffices to show that $\tilde{\mathbb{P}}((\tilde{X}_n, \tilde{Y}_n) \in A) = \mathbb{P}^n(A)$ for each A such that $\mathbb{P}^n(\partial A) = 0$. Let A be such a set. Let A^k be the union of the sets $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k}$ such that $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k} \subseteq A$ and let A'^k be the union of the sets

$E^1_{i_1,\ldots,i_k} \times E^2_{i'_1,\ldots,i'_k}$ such that $E^1_{i_1,\ldots,i_k} \times E^2_{i'_1,\ldots,i'_k} \cap A \neq \emptyset$. Then $A^k \subseteq A \subseteq A'^k$ and, by the construction above,

$$\tilde{\mathbb{P}}((\tilde{X}_n, \tilde{Y}_n) \in A^k) = \mathbb{P}^n(A^k) \text{ and } \tilde{\mathbb{P}}((\tilde{X}_n, \tilde{Y}_n) \in A'^k) = \mathbb{P}^n(A'^k)$$

Now let $C^k = \{s \in \mathbf{E}_1 \times \mathbf{E}_2 : \pi(s, \partial A) \leq 2^{-k}\}$. Then $A'^k - A^k \downarrow \partial A$ as $k \to \infty$. Since $\mathbb{P}^n(\partial A) = 0$ by assumption, $\mathbb{P}^n(C^k) \downarrow 0$ as $k \to \infty$. Hence

$$\tilde{\mathbb{P}}((\tilde{X}_n, \tilde{Y}_n) \in A) = \lim_{k\to\infty} \tilde{\mathbb{P}}((\tilde{X}_n, \tilde{Y}_n) \in A^k) = \lim_{k\to\infty} \mathbb{P}^n(A^k) = \mathbb{P}^n(A).$$

Following the same way, we can show that $\tilde{X}$ has probability law $\mathbb{P}_1$. □

REFERENCES

[1] Aksin, Z., Armony, M., and Mehrotra, V., "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, no. 6, pp. 665 – 688, 2007.

[2] Asmussen, S., *Applied probability and queues*, vol. 51 of *Applications of Mathematics (New York)*. New York: Springer-Verlag, second ed., 2003.

[3] Avi-Itzhak, B. and Halfin, S., "Expected response times in a non-symmetric time sharing queue with a limited number of service positions," in *Proceedings of the 12th International Teletraffic Congress*, (Torino), 1988.

[4] Billingsley, P., *Probability and measure*. Wiley Series in Probability and Mathematical Statistics, New York: John Wiley & Sons Inc., third ed., 1995.

[5] Billingsley, P., *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics, New York: John Wiley & Sons Inc., second ed., 1999.

[6] Blake, R., "Optimal control of thrashing," in *Proceedings of the 1982 ACM SIGMETRICS Conference on Measurements and Modeling of Computer Systems*, (Seattle, WA), 1982.

[7] Bramson, M., "State space collapse with application to heavy traffic limits for multiclass queueing networks," *Queueing Systems Theory Appl.*, vol. 30, no. 1-2, pp. 89–148, 1998.

[8] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., "Statistical analysis of a telephone call center: a queueing-science perspective," *J. Amer. Statist. Assoc.*, vol. 100, no. 469, pp. 36–50, 2005.

[9] Budhiraja, A. and Lee, C., "Stationary distribution convergence for generalized Jackson networks in heavy traffic," tech. rep., University of North Carolina at Chapel Hill, 2008.

[10] Chung, K. L., *A course in probability theory*. San Diego, CA: Academic Press Inc., third ed., 2001.

[11] Dai, J. G., "On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models," *Ann. Appl. Probab.*, vol. 5, no. 1, pp. 49–77, 1995.

[12] Dai, J. G., "Stability of open multiclass queueing networks via fluid models," in *Proceedings of the IMA workshop on stochastic networks*, (New York), Springer-Verlag, 1995.

[13] Dai, J. G., He, S., and Tezcan, T., "Many-server diffusion limits for $G/Ph/n + GI$ queues," tech. rep., Georgia Institute of Technology, 2009.

[14] Denning, P. J., Kahn, K. C., Leroudier, J., Potier, D., and Suri, R., "Optimal multiprogramming," *Acta Informatica*, vol. 7, pp. 197–216, 1976.

[15] Doytchinov, B., Lehoczky, J., and Shreve, S., "Real-time queues in heavy traffic with earliest-deadline-first queue discipline," *Ann. Appl. Probab.*, vol. 11, no. 2, pp. 332–378, 2001.

[16] Elnikety, S., Nahum, E., Tracy, J., and Zwaenepoel, W., "A method for transparent admission control and request scheduling in e-commerce web sites," in *World-Wide-Web Conference*, 2004.

[17] Ethier, S. N. and Kurtz, T. G., *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, New York: John Wiley & Sons Inc., 1986.

[18] Gamarnik, D. and Momčilović, P., "Steady-state analysis of a multi-server queue in the halfin-whitt regime," 2007.

[19] Gamarnik, D. and Zeevi, A., "Validity of heavy traffic steady-state approximation in generalized Jackson networks," *Ann. Appl. Probab.*, vol. 16, no. 1, pp. 56–90, 2006.

[20] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers:tutorial,review, and research prospects.," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[21] Garnet, O., Mandelbaum, A., and Reiman, M., "Designing a call center with impatient customers," *Manufacturing & Service Operations Management*, vol. 4, no. 3, pp. 208–227, 2002.

[22] Grishechkin, S., "$GI/G/1$ processor sharing queue in heavy traffic," *Adv. in Appl. Probab.*, vol. 26, no. 2, pp. 539–555, 1994.

[23] Gromoll, H. C., "Diffusion approximation for a processor sharing queue in heavy traffic," *Ann. Appl. Probab.*, vol. 14, no. 2, pp. 555–611, 2004.

[24] Gromoll, H. C. and Kruk, Ł., "Heavy traffic limit for a processor sharing queue with soft deadlines," *Ann. Appl. Probab.*, vol. 17, no. 3, pp. 1049–1101, 2007.

[25] Gromoll, H. C., Puha, A. L., and Williams, R. J., "The fluid limit of a heavily loaded processor sharing queue," *Ann. Appl. Probab.*, vol. 12, no. 3, pp. 797–859, 2002.

[26] Gromoll, H. C., Robert, P., and Zwart, B., "Fluid limits for processor sharing queues with impatience," *Math. Oper. Res.*, 2008. to appear.

[27] Gupta, V., Dai, J. G., Harchol-Balter, M., and Zwart, B., "On the inapproximability of $M/G/K$: Why two moments of job size distribution are not enough," tech. rep., Carnegie Mellon University, 2007.

[28] Halfin, S. and Whitt, W., "Heavy-traffic limits for queues with many exponential servers," *Oper. Res.*, vol. 29, no. 3, pp. 567–588, 1981.

[29] Heiss, H.-U. and Wagner, R., "Adaptive load control in transaction processing systems," in *Proceedings of the 17th International Conference on Large Data Bases*, 1991.

[30] Hunter, J. K. and Nachtergaele, B., *Applied analysis*. River Edge, NJ: World Scientific Publishing Co. Inc., 2001.

[31] Jean-Marie, A. and Robert, P., "On the transient behavior of the processor sharing queue," *Queueing Systems Theory Appl.*, vol. 17, no. 1-2, pp. 129–136, 1994.

[32] Jelenković, P., Mandelbaum, A., and Momčilović, P., "Heavy traffic limits for queues with many deterministic servers," *Queueing Syst. Theory Appl.*, vol. 47, no. 1/2, pp. 53–69, 2004.

[33] Kallenberg, O., *Random measures*. Berlin: Akademie-Verlag, fourth ed., 1986.

[34] Kamra, A., Misra, V., and Nahum, E. M., "Yaksha: A self-tuning controller for managing the performance of 3-tiered web sites," in *Twelfth IEEE International Workshop on Quality of Service*, 2004.

[35] Kang, K. and Ramanan, K., "Fluid limits of many-server queues with reneging," tech. rep., Carnegie Mellon University, 2008.

[36] Kaspi, H. and Ramanan, K., "Law of large number limits for many-server queues," 2007. working paper.

[37] Kleinrock, L., *Queueing systems*. New York: Wiley-Interscience, 1976. Computer Applications, Volume II.

[38] Lang, S., *Real analysis*. Reading, MA: Addison-Wesley Publishing Company Advanced Book Program, second ed., 1983.

[39] Mandelbaum, A. and Momčilović, P., "Queues with many servers: The virtual waiting-time process in the qed regime," *Mathematics of Operations Research*, vol. 33, no. 3, pp. 561–586, 2008.

[40] Mandelbaum, A. and Momčilović, P., "Queues with many servers and impatient customers," 2009.

[41] Mandelbaum, A. and Shimkin, N., "A model for rational abandonments from invisible queues," *Queueing Systems Theory Appl.*, vol. 36, no. 1-3, pp. 141–173, 2000.

[42] Nuyens, M. and van der Weij, W., "The limited processor sharing queue," tech. rep., CWI, Amsterdam, 2007.

[43] Pang, G. and Whitt, W., "Two-parameter heavy-traffic limits for infinite-server queues," tech. rep., Columbia University, 2008.

[44] Puha, A. L., Stolyar, A. L., and Williams, R. J., "The fluid limit of an overloaded processor sharing queue," *Math. Oper. Res.*, vol. 31, no. 2, pp. 316–350, 2006.

[45] Puha, A. L. and Williams, R. J., "Invariant states and rates of convergence for a critical fluid model of a processor sharing queue," *Annals of Applied Probability*, vol. 14, no. 2, pp. 517–554, 2004.

[46] Puhalskii, A. A. and Reed, J. E., "On many-server queues in heavy traffic," *Annals of Applied Probability*, 2009. To appear.

[47] Puhalskii, A. A. and Reiman, M. I., "The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime," *Adv. in Appl. Probab.*, vol. 32, no. 2, pp. 564–595, 2000.

[48] Reed, J. E., "The $G/GI/N$ queue in the Halfin-Whitt regime," tech. rep., Georgia Institute of Technology, 2007.

[49] Ritchie, D. M. and Thompson, K., "The Unix time-sharing system," *J. ACM*, vol. 17, no. 7, pp. 365–375, 1974.

[50] Schroeder, B., Harchol-Balter, M., Iyengar, A., Nahum, E., and Wierman, A., "How to determine a good multi-programming level for external scheduling," in *Proceedings of the 22nd International Conference on Data Engineering*, (Atlanta, GA), April 2006.

[51] Sigman, K. and Wolff, R. W., "A review of regenerative processes," *SIAM Rev.*, vol. 35, no. 2, pp. 269–288, 1993.

[52] Talreja, R. and Reed, J. E., "Distribution-valued heavy-traffic limits for the g/gi/infinity queue," tech. rep., Columbia University, 2008.

[53] van der Vaart, A. W. and Wellner, J. A., *Weak convergence and empirical processes*. Springer Series in Statistics, New York: Springer-Verlag, 1996.

[54] Whitt, W., *Stochastic-process limits*. Springer Series in Operations Research, New York: Springer-Verlag, 2002. An introduction to stochastic-process limits and their application to queues.

[55] Whitt, W., "Efficiency-driven heavy-traffic approximations for many-server queues with abandonments," *Mgt. Sci.*, vol. 50, no. 10, pp. 1449–1461, 2004.

[56] Whitt, W., "Fluid models for multiserver queues with abandonments," *Oper. Res.*, vol. 54, no. 1, pp. 37–54, 2006.

[57] Williams, R. J., "Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse," *Queueing Systems Theory Appl.*, vol. 30, no. 1-2, pp. 27–88, 1998.

[58] Zeltyn, S. and Mandelbaum, A., "Call centers with impatient customers: Many-server asymptotics of the $M/M/n + G$ queue," *Queueing Syst. Theory Appl.*, vol. 51, no. 3-4, pp. 361–402, 2005.

[59] Zhang, F. and Lipsky, L., "Modelling restricted processor sharing," in *Proc. of the 2006 Int'l Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA06)*, 2006.

[60] Zhang, F. and Lipsky, L., "An analytical model for computer systems with non-exponential service times and memory thrashing overhead," in *Proc. of the 2007 Int'l Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA07)*, 2007.

[61] Zhang, J., "Fluid models of multi-server queues with abandonment," tech. rep., Georgia Institute of Technology, 2008.

[62] Zhang, J., Dai, J. G., and Zwart, B., "Diffusion limits of limited processor sharing queues," tech. rep., Georgia Institute of Technology, 2007.

[63] Zhang, J., Dai, J. G., and Zwart, B., "Law of large number limits of limited processor sharing queues," *Math. Oper. Res.*, 2009. To appear.

[64] Zhang, J. and Zwart, B., "Steady state approximations of limited processor sharing queues in heavy traffic," *Queueing Systems. Theory and Applications*, vol. 60, no. 3-4, pp. 227–246, 2008.